

Index

Numerics

“3x higher churn”, 247, 248

A

account takeover fraud, 300

accuracy metrics, 118, 124, 126, 128–129

action plan, data science project, 356

Advanced Data Analysis (ADA), 49, 61–63, 87

 CSV to ChatGPT, 18

adversarial adaptation, 304

AI. *See* artificial intelligence (AI)

alternatives analysis

 clear and reasonable set, 358

 criteria for, 358

 documentation, 358

 evaluations, 358–359

 data resources, 359

 data skillset, 358

 data technologies, 358

 project objectives, 358

analyst, segmentation, 168–170

analytical dashboards, 198

analytical work, connects to financial outcomes,
 244–246

 business impact, 246

 decision-making, 244

 executives and stakeholders’ language, 245

 profit, loss, and operational efficiency, 244–245

analytics leader path, 364

analytics power user prompts, 383–384

 auditing data quality, 383–384

 blueprinting a dashboard, 387

 converting an analysis into a report, 388–389

 debugging an incorrect analysis, 389–390

 decoding AI-generated code, 391

 designing a visualization, 386–387

 interpreting statistical results, 387–388

 kicking off an exploratory analysis, 384–385

 setting up a predictive model, 385–386

 translating a technical concept, 390

annotations, context creation with, 231–232

anomalies categories, EDA, 101

anomaly detection, 301–302

 definition, 301

 supervised, 301, 302

 unsupervised, 301, 302

AOV. *See* average order value (AOV)

application fraud, 301

Application Programming Interface (API), 40–42,
 63, 67, 73

area charts, 224

artificial intelligence (AI), 2

 assistance, 41–42

 -assisted approach, 20–22, 255

 coding skills, 21–22

 comparison with old way vs,
 16–17

 Python toolkit, 21

 -assisted dynamic pricing, 267

 -assisted lead scoring, 281

 business processes with, 255

 categories data, 25–26

 category labels, 25–26

 copilots, 64, 194, 283–284, 286

 in modern BI tools, 197

 for correlation, 154

 data cleaning. *See* data cleaning

 data quality

 duplicate records, 32–33

 formatting inconsistencies,
 33–34

 missing values, 31

 outlier value (signal/error), 32

 data science. *See* data science, AI

 dataset, 35–36

 data types

 categories and labels, 25–26

 dates and times, 26

 number types, 24–25

 text data, 27

 -driven data job, 367

 interview preparation, 369

 networking in an authentic manner, 369

 personal brand, 368

 portfolio, 367–368

 duplicate removal with, 47

artificial intelligence (AI) (*continued*)

- ethics assessments, 340–345

- accountability of, 342

- bias evaluation, 344–345

- trustworthy AI, 341

- unbiased results, 342–344

- ethics frameworks, 352

- Exploratory Data Analysis (EDA)

- business model, 100–101

- categories of discovery, 101

- distribution shapes, 102–103

- hypotheses to test, 101–102

- identifying relationships between variables, 104–106

- outliers, 104

- pandas, 114–115

- patterns and clusters, 106

- questions, 106–108

- seaborn, 115

- systematic exploration with Claude, 108–114

- threads, 108

- to flag interesting patterns, 113–114

- generated budget recommendation, 277–278

- generated charts, 222

- generated code, decoding, 391

- machine learning (ML) model, 117–120,

- 124–128, 135–136

- accuracy metrics, 128–129

- decision tree, 125–126

- explaining, 126–128

- memorizes, 131–132

- supervised learning, 121–124

- training/testing accuracy, 130

- trust vs escalate decision framework, 134–135

- underfitting, 132–133

- operational context, 258–259

- plain-English interpretations, 37

- product manager, 365

- prompt engineering

- adapting, 87–88

- analysis formats, 82–83

- churn analysis, 77, 93, 94, 97

- code, 97–98

- complex questions into simple, 80–82

- context and constraints, 83

- documenting successful prompts for reuse, 86–87

- interprets your requests, 79

- prompt iteration, 79–80

- prompt templates, 84–85, 91–94

- refinement approach, 95–96

- results, 83–84

- tools, 87–88

- vague vs specific, 78–79

- workspace in ChatGPT Projects, 88–91

- prompting, 257

- logistics cost-savings demonstration, 258–261

- operationally fluent prompting, 257–258

- search, 270

- statistics, 27

- comparison with mean, median and mode, 27–28

- intuition, 38

- spotting relationships using correlation, 30–31

- spread and variability, 28–30

- step-by-step breakdown, 36–37

- tools, 1, 3–5, 257

- comparison with ChatGPT, Claude, and Gemini, 62–63

- data categories, 73–74

- data minimization, 73

- failure modes, 65–67

- fast verification techniques, 67–70

- multi-tool approach, 64

- practical mental models, 60–62

- pricing models, 72–73

- right tool for different tasks, 63–64

- safer workflows, 74

- trust rubric, 70–72

- uses probability for predictions, 143–144

- Ask Data panel, 396

- assumption errors, 66

- attrition risk, 292

- auditing data quality, 383–384

- average order value (AOV), 147

B

- bar charts, 197, 224, 225, 228, 229, 373, 396

- bell-shaped (normal) distributions, 103

- BI. *See* business intelligence (BI)

- BigQuery, 42, 63, 64, 196, 286

- Google BigQuery with Gemini, 397

- bimodal distributions, 103

- black box

- approach, 343

- warning, 341

- blueprinting a dashboard, 387

- brainstorming, 216–218

- breakdown, of template framework, 91

- bubble plots, 225

- budget planning, forecasts for, 186–187

- business acumen, 239

- business conversations, 246–247

- add human context layer, 248–249

- AI to draft three versions, 247–248

- stakeholders mapping, 247

- business landscape traverse, 240–244

- data work connection, 244–246

- business impact, 246
- decision-making, 244
- executives and stakeholders' language, 245
- profit, loss, and operational efficiency, 244–245
- defined, 239
- executive mentality, 239
- financial savvy, 239
- layers of, 242
- leadership skills, 239
- leveling up, 242–244
- with subject matter expertise, 238–239
- business applications, regression for, 174
- business context, 16, 33, 57, 68, 79, 83, 89, 92
- business conversations, 246–247
 - add human context layer, 248–249
 - AI to draft three versions, 247–248
 - stakeholders map, 247
- business decisions, turning forecasts into, 186–188
 - communicating to stakeholders, 187–188
 - forecasts for budget planning, 186–187
- business gap, 237–239
- business intelligence (BI), 283–285
 - copilots add value, 287
 - ad hoc questions, 287
 - executive-level self-service, 287
 - iterative question-asking, 287
 - decision-support products, 285
 - evolved with AI assistance, 285–286
 - historical, static data, 285
 - internal data, 285
 - natural-language querying, 286–287
 - small-and medium-size datasets, 285–286
 - tool, 194
 - assessing the feature alternatives, 196–197
 - examining cost structure, 194–195
 - integrating with existing ecosystems, 196
 - learning curves and AI assistance, 195–196
 - leveraging AI copilots in modern, 197
 - selection, 194–197
 - technologies, and skillsets, 286
- business landscape traverse, 240–244
 - cost savings, 240
 - developing business acumen, 242–244
 - revenue-generation, 240
 - traditional data roles, 240–242
- business process automation (BPA), 255
 - AI-assisted approach, 255
 - cycle time analysis, 255
 - demand and capacity forecasting, 255
 - exception detection and escalation routing, 255
 - process intelligence and workflow mining, 255

- business profit, 239, 240, 242, 284, 285
- business vision, 314–316

C

- California Consumer Privacy Act (CCPA), 273, 352
- campaign performance data, 274–275
- card testing fraud, 299, 300, 307, 308
- catching calculation, 65–66
- causation vs correlation, 139–141
- CCPA. *See* California Consumer Privacy Act (CCPA)
- centroid, 159, 161
- CEO, 173, 247, 248, 383
- channel
 - defined, 270
 - mapping, 269–270
 - performance, 270–271
 - scorecard, 273
- channel-by-channel basis, 269
- channel scoring, 267, 271–273
 - customer avatar, 272
 - evaluate metrics, 271–272
 - mapping, 271
 - sales and marketing strategy, 272
 - scorecard creation, 272
- ChatGPT, 2–5, 9, 11–13, 15, 17, 21, 39, 41, 49, 50, 52, 54, 55, 77, 78, 87, 93, 100, 108, 118, 125–128, 130–132, 135, 136, 272
- Advanced Data Analysis (ADA), 18, 124, 176, 394
 - uploading marketing data to, 176
- chart types, 94–95
- cleaning data with, 34–35
- clustering customers with, 163–168
- clusters, 170–171
- comparison with Claude, Gemini and, 62–63
- cost-savings analysis, 261
 - AI recommendation, 262
 - impact and feasibility, 263
 - operations datasets, 262
- to create a sensitivity analysis, 185
- data quality assessment by, 19–20
- marketing spend recommendations from, 273–274
 - campaign performance data, 274–275
 - reallocation, 276–277
 - testing AI suggestions, 277–278
 - writing fluent prompt, 275–276
- risk and fraud detection, 305–309
 - building processes, 308–309
 - flag suspicious activities, 306–307
 - investigating flagged transactions, 307–308
 - transaction data, 306

- ChatGPT Projects, 88–91
 - workspace in, 88–91
 - analytics, 89
 - approach, 90–91
 - requirements, 90
 - uploading and organizing datasets, 89–90
- Choropleth, 228, 229
- class imbalance, 304
- Claude, 2–5, 77, 78, 87–89, 100
 - by anthropic, 394–395
 - comparison with ChatGPT, Gemini and, 62–63
 - systematic exploration with, 108
 - AI to flag interesting patterns, 113–114
 - dataset, 109–110
 - distribution analysis and visualizations, 110–111
 - identifying correlations and relationships, 111–113
- cleaning data. *See also* data cleaning
 - with ChatGPT, 34–35
 - with pandas library, 54–56
 - with SQL scripts, 56–57
- clustering, 158
 - algorithms, 159
 - current segmentation feels crude, 158
 - customers with ChatGPT, 163–168
 - k-means, 160
 - k value selection, 162–163
 - methodology, 268, 279
 - unexpected patterns, 159
- clusters, ChatGPT, 170–171
- coding skills, 14, 21–22
- collaborative or content-based filtering, 266
- color and clarity principles, 221–222
- comma-separated values (CSV) file, 39–41, 49, 57, 109
 - to ChatGPT ADA, 18
- communication to stakeholders, 187–188
- company information gathering
 - data ethics, 327–329
 - AI ethics assessment, 328–329
 - data privacy rights, 328
 - data resources, 319–324
 - data silo analysis, 322–324
 - dictionary request, 320–321
 - inventory request, 320–321
 - surveys and interviews files, 321–322
- data science framework, 315–317
- data science project pitfalls, 325–327
 - best practices, 326–327
 - business-focused strategy, 326
- data technologies, 317–319
 - data architecture, 317, 318
 - surveys and interviews, 318, 319
- efficiency framework, 329–331
- IT technology infrastructure, 317
- people-mapping, 324–325
 - data skillsets, 324–325
 - organizational charts, 324
- revenue/business vision, 314–315
- vision, mission, and values, 315–317
- company-specific expertise, 243
- company vision, mission, and values, 315–317
- comparative graphics, 225–226
 - bubble plots, 225
 - Gantt charts, 225
 - packed circle diagrams, 225
 - stacked charts, 225
 - treemaps, 225
 - word clouds, 225
- complex questions, 77, 80–82, 107–108
- confidence interval, 175
- confidential business information, 73
- constraints, 83, 92
 - framing, 258
 - of template framework, 91–92
- content creation, 266–267
- context
 - creation
 - with annotations, 231–232
 - with data, 231
 - with graphical elements, 232–233
 - of template framework, 91–92
- Conway, Drew, 238
- Copilot, Microsoft Power BI with, 396
- correlation
 - AI for, 154
 - analysis by Claude, 112
 - vs causation, 139–141
 - hunting, 108
 - identifying relationships and, 111–113
 - matrices, 115
 - measurement, 138
 - Pearson, 154
 - and relationships, 138–141
 - spotting relationships using, 30–31
 - variables relate to each other, 138–139
- correlation coefficients, 137
- cost savings, 240, 254
 - demonstration, 258–261
- cost-savings analysis, ChatGPT, 261
 - AI recommendation, 262
 - impact and feasibility, 263
 - operations datasets, 262
- cross-validation, 69
- C-suite, 188, 198, 239, 245

- customer churn, 266
 - analysis, 266
 - reduction, 266
- customer segmentation, 278–280
- cycle time analysis, 255

D

- dashboard
 - analytical, 198
 - blueprinting, 387
 - design fundamentals, 197
 - balancing detail with clarity, 200–201
 - designing for audience, 198–199
 - organizing information hierarchically, 200
 - right metrics to display, 199
 - vibe-coding a dashboard, 201–202
 - detailed breakdowns, 200
 - executive, 198
 - KPI headlines, 200
 - operational, 198
 - supporting trends, 200
 - vibe-coding dashboard in lovable, 201–202
 - building a dashboard, 203–209
 - examining the dataset, 202–203
- dashboards with AI, 193–194
 - coding up dashboards with Plotly, 209–211
 - dashboard design fundamentals, 197–202
 - right BI tool selection, 194–197
 - assessing the feature alternatives, 196–197
 - considering learning curves and AI assistance, 195–196
 - examining cost structure, 194–195
 - integrating with existing ecosystems, 196
 - leveraging AI copilots in modern, 197
- vibe-coding first dashboard in lovable, 202–209

- data analysis, Exploratory Data Analysis (EDA)
- business model, 100–101
- categories of discovery, 101
- distribution shapes, 102–103
- hypotheses to test, 101–102
- identifying relationships between variables, 104–106
- outliers, 104
- pandas, 114–115
- patterns and clusters, 106
- questions, 106–108
- seaborn, 115
- systematic exploration with Claude, 108–114
- threads, 108
- data analytics, 288
- challenges in, 289
- data wrangling, 289–290

- to support decision-making, 287–289
 - descriptive analytics, 288
 - diagnostic analytics, 288
 - predictive analytics, 288
 - prescriptive analytics, 288
- data and AI products building, 278–279
 - customer segmentation, 279–280
 - lead scoring systems, 280–281
- data architecture, 290, 318
- data art, 214, 218, 220
 - for activists, 215–216
- DataCamp and Kaggle for Learning, 399–400
- data cleaning, 15, 16, 48–49
 - categorical data and text cleaning, 45–47
 - date formats, 50
 - dealing with outliers and errors, 47–48
 - delete rows with missing values, 43
 - duplicate removal with AI, 47
 - exporting, 54
 - fill missing values, 43
 - finding and importing, 40–42
 - flag missing values, 44
 - identifying and removing duplicates, 51–52
 - initial assessment, 49–50
 - investigating outliers, 52–53
 - missing values, 50–51
 - with pandas library, 54–56
 - SQL scripts, 56–57
 - standardizing categories, 51
 - standardizing dates and times, 44–45
 - validating, 53
- data context, 79, 83
- data dictionary, 320–321, 337
 - vs dataset, 320, 321
 - structure, 320
- data errors, 101, 104, 149, 150
 - outliers, 150
- data ethics, 327–329
 - AI ethics assessment, 328–329
 - data privacy rights, 328
- data extraction, 40, 289
- data generation activities, 353
- data governance, 290, 345–347
 - council, 346, 347
 - standards, 329, 347
- data graphics, 229–230
 - adding context, 231
 - context creation
 - with annotations, 231–232
 - with data, 231
 - with graphical elements, 232–233
- data visualizations vs, 222

- data graphics (*continued*)
 - type selection, 222–223
 - comparative graphics, 225–226
 - spatial plots and maps, 228–229
 - standard chart graphics, 224–225
 - statistical plots, 226–227
 - topology structures, 227–228
- data literacy, 23, 244, 289, 326, 328, 331, 343, 355
- data minimization, 73, 347
- data mining, 289, 290
- data points, outliers and unusual, 149–151
- data preparation, 39, 40, 290
- data privacy, 306, 342, 343
 - laws, 352
 - policy, 345–347
- data products, 268, 278–280
- data professionals, 239, 240, 243, 245, 246, 249
- data program status and recommendations, 354–356
 - business requirements and needs, 354
 - data initiative best practices, 356
- data resources, 354–355
 - data governance, 355
 - legal, regulatory, and AI ethics, 354
- data teams, 356
- data technology, 354
 - reference architecture, 354, 355
 - summary, 354
- goals, 354
 - business mission goals, 354
 - data science project goals, 354
- new hires, 356
- personnel repositioning, 356
- skillset and people, 355
 - data literacy training, 355
 - data team personnel training, 355
 - use case skillset requirements, 355
 - user training, 355
- data quality, 15, 17, 18
 - assessment by ChatGPT, 19–21
 - auditing, 383–384
 - duplicate records, 32–33
 - formatting inconsistencies, 33–34
 - missing values, 31
 - outlier value (signal/error), 32
- data resources, 319–324
 - data dictionary, 320–321
 - vs dataset, 320, 321
 - structure, 320
 - data generation activities, 353
 - data inventory, 320–321
 - data security processes, 353
 - recommendations, 354–355
 - data governance, 355
 - legal, regulatory, and AI ethics, 354
 - silos analysis, 322–324
 - surveys and interviews files, 321–322
- data roles and business metrics, 241
- data science
 - framework, 315–317
 - Venn diagram, 238
- data science, AI
 - AI-assisted approaches, 20–22
 - coding skills, 21–22
 - Python toolkit, 21
 - ChatGPT, 12–13
 - comparison with old way vs AI-assisted way, 16–17
 - definition, 14–15
 - generative AI (Gen AI), 9–12, 17–19
 - CSV to ChatGPT ADA, 18
 - data quality assessment by ChatGPT, 19–20
 - reshaping data analysis, 15–16
 - tools for analytical tasks, 13–14
 - large language models (LLMs), 11–12, 14
 - machine learning, 11
 - rapid-fire change, 10–11
- data science career matrix, navigation, 364
 - AI product manager path, 365
 - analytics leader path, 364
 - machine learning (ML) engineer path, 366
- data science leader, 242, 243, 314
- data science leadership, 21
- data science project pitfalls, 325–327
 - best practices, 326–327
 - business-focused strategy, 326
- data science project plan
 - execution phase, 362
 - implementation plan, 350–357
 - action plan, 356
 - alternative solutions, 354
 - data program status, 354–356
 - data resources, 353
 - data skillsets, 353
 - data technologies, 353
 - legal, regulatory, and ethical framework, 352
 - organizational structure, 352–353
 - organization overview, 352
 - recommendations, 354–356
 - STAR framework, 350–351
 - summary statement, 351
 - template, 351–357
 - purpose of, 350

- supports, 357–362
 - alternatives analysis, 358–359
 - processes, organization, technology, information (POTI) model, 360–362
 - users interview, 359–360
- data science use cases
- documentation review, 334
- plug-and-play assessments, 338–347
 - AI ethics assessment, 340–345
 - data governance assessment, 345–347
 - data privacy policy assessment, 345–347
 - data skills gap analysis, 338–340
- quick win, 335–338
 - business capability assessment, 335–336
 - processes, organization, technology, information (POTI) model, 336–338
 - selection, 335–336
 - STAR framework, 334, 335, 338
- data scientists, 15, 20, 99, 216, 238, 241, 288, 314, 343
- data security processes, 353
- datasets
 - practice, 35–36
 - uploading and organizing, 89–90, 109–110
- data showcasing for analysts, 215
- data skillsets, 336, 338, 339. *See also* skillset and people recommendations
 - inventory, 353
 - methods, 353
- data skills gap analysis, 338–340
- data storytelling, 218, 219
 - for decision makers, 214–215
- data teams, 12, 314, 353, 354, 356
- data technology, 317–319, 353, 354, 358
 - data architecture, 317, 318
 - recommendations
 - reference architecture, 354, 355
 - summary, 354
 - surveys and interviews, 318, 319
- data transformation, 15–16
- data types
 - categories and labels, 25–26
 - dates and times, 26
 - number types
 - decimals, 24
 - integers, 24
 - percentages, 24–25
 - text data, 27
 - and unit mistakes, 66–67
- data visualizations, 95, 110–111, 213, 214, 222, 228, 230
 - adding context, 231
 - color and clarity principles, 221–222
 - context creation
 - with data, 231
 - with graphical elements, 232–233
- data art, 215–216
- vs data graphics, 222
- data showcasing, 215
- data storytelling, 214–215
- design style, 219
 - calculating and exacting feel, 219–220
 - emotional response, 220–221
- persuasive, 233
- pitfalls in, 228
- skills, 218
- standard chart graphics, 224–225
- data work connection to revenue and strategy, 244–246
 - business impact, 246
 - decision-making, 244
 - executives and stakeholders' language, 245
 - profit, loss, and operational efficiency, 244–245
- data wrangling, 289–290
 - data architecture, 290
 - data extraction, 289
 - data governance, 290
 - data preparation, 290
- date formats, 26
 - fixing, 50
 - standardizing, 44–45
- decimals, types of numbers, 24
- decision anchor, 258
- decision-making, 218, 284
- decision-support products, 284, 285
- decision-support systems, 284
- decision tree, 125–126, 288, 291, 292
- decoding AI-generated code, 391
- decomposition, time series, 153
- delete rows, with missing values, 43
- Delivery Logistics dataset, 258
- demand and capacity forecasting, 255
- .describe(), 114–115
- descriptive analysis, vs predictive analytics, 119
- descriptive analytics, 285, 288
- design artistry, 220
- design style, 219
 - calculating and exacting feel, 219–220
 - emotional response, 220–221
- diagnostic analytics, 288
- distribution
 - analysis, 110–111
 - shapes, 102–103
- documentation, Notion AI for, 398
- domain declaration, 257
- domain-specific pattern analysis, 259–260

duplicate records, 32–33
dynamic pricing, 267

E

e-commerce businesses, 163, 303, 304
EDA. *See* Exploratory Data Analysis (EDA)
edge cases, 294
efficiency gains, opportunities for, 253
elbow plot, 162
emotionally provocative data visualizations, 220–221
energy consumption pattern analysis, 256
entry-level analysts, 364
equipment sensor, 254
equivalent SQL (structured query language) query, 292
errors, 32
 assumption, 66
 data, 101, 104, 149, 150
 dealing with, 47–48
 logic, 65–66
 Mean Absolute Error (MAE), 146, 181, 182
 Root Mean Squared Error (RMSE), 146, 180–182
 sanity checks catch, 68
escalation routing, 255
Excel (XLSX) file, 40–41
exception detection, 255
executive dashboards, 198, 205, 207, 211
executive mentality, 239
executives' language, 245
exploratory analysis, 244, 384–385
Exploratory Data Analysis (EDA), 119, 149
 business model, 100–101
 categories of discovery, 101
 distribution shapes, 102–103
 hypotheses to test, 101–102
 identifying relationships between variables, 104–106
 outliers, 104
 pandas, 114–115
 patterns and clusters, 106
 questions, 106–108
 seaborn, 115
 systematic exploration with Claude, 108–114
 threads, 108
extrapolation risk, 184
eyeball method, 32

F

failure modes, for AI tools, 65–67
 assumption errors, 66
 catching calculation and logic errors, 65–66

 data-type and unit mistakes, 66–67
 hallucination patterns, 65
 workflow failures, 67
false negative, 301–303
false positive, 301–303, 305
fast verification techniques, 67–70
 authoritative sources, 69
 critical numbers, 68
 cross-validation, 69
 documenting discrepancies and corrections, 69–70
 sanity checks catch errors, 68
feasibility, 263
fill missing values, 43
financial savvy, 239
first-pass regression model, 177–178
`fit_predict()`, 171
fit regression, 137
flag interesting patterns, AI, 113–114
flag missing values, 44
floats/real numbers, 24
fluent prompt, writing, 275–276
forecasting, time series analysis, 153
format specification, of template framework, 91–92
formatting inconsistencies, 33–34
fraud
 account takeover, 300
 application, 301
 card testing, 299, 300, 307, 308
 identity, 301
 novel fraud schemes, 304
 velocity, 300–301
fraud detection, risk and
 anomaly detection, 301–302
 definition, 301
 supervised anomaly detection, 301, 302
 unsupervised anomaly detection, 301, 302
 with ChatGPT, 305–309
 building processes, 308–309
 flag suspicious activities, 306–307
 investigating flagged transactions, 307–308
 transaction data, 306
e-commerce and financial services, 300
false negative, 302–303
false positive, 302–303
with generative AI, 303–305
 with human judgment, 305
 limitations, 304–305
 unusual patterns, 303–304
overview, 300–303
pattern recognition, 300–301
 account takeover fraud, 300
 application fraud, 301

- card testing, 300
- identity fraud, 301
- velocity fraud, 300–301

functional visualization, 216, 218–219

G

Gantt charts, 225

“garbage in, garbage out” principle, 175

GDPR. *See* General Data Protection Regulation (GDPR)

Gemini, 2, 3, 11, 34, 42, 60, 63–64, 77, 78, 87, 88

- comparison with ChatGPT, Claude, and 62–63
- Google BigQuery with, 397
- in Looker, 196

Gen AI. *See* Generative AI (Gen AI)

general business expertise, 242

General Data Protection Regulation (GDPR), 326, 342, 343, 352

- Article 83(4), 343
- Article 83(5), 343
- data privacy rights, 343
- Recital 71, 343
- violation of, 343

Generative AI (Gen AI), 1, 9–12, 17–19, 363

- CSV to ChatGPT ADA, 18
- data quality assessment by ChatGPT, 19–20
- practical mental models, 60–61
- reshaping data analysis, 15–16
- risk and fraud detection, 303–305
 - with human judgment, 305
 - limitations, 304–305
 - unusual patterns, 303–304
- tools for analytical tasks, 13–14

generic prompts, 257, 274

Goldilocks principle, 133

“Good fit”, 180–182

Google AI Pro, 395

Google Analytics (GA4), 271

Google BigQuery with Gemini, 397

Google Cloud, 397

Google Cloud Platform commitments, 195

Google Gemini, 395

GPT-3-era tools, 266

granularity, 290

graphical elements, context creation with, 232–233

graph models, 227

gross margin, 245, 256

gross profit, 245

GROUP BY, 297

groupby(‘segment’).mean(), 171

H

hallucination patterns, 12, 65

high-stakes decisions, 70, 134, 176

high-trust scenarios, 71

histogram, 110, 111, 226

human context layer, 248–249

hyper-targeted advertising, 267

hypotheses, 101–102

I

IBM HR Analytics Employee Attrition, 291

IBM HR Analytics Employee Attrition & Performance, 124

identity fraud, 301

impact vs feasibility, AI-generated cost-saving suggestions, 263

implementation plan, 350–357

- action plan, 356
- alternative solutions, 354
- data program status, 354–356
- data resources, 353
- data skillsets, 353
- data technologies, 353
- legal, regulatory, and ethical framework, 352
- organizational structure, 352–353
- organization overview, 352
- recommendations, 354–356
- STAR framework, 350–351
- summary statement, 351
- template, 351–357

implementation roles call, 239

industrial machine learning, manufacturing operations

- defect analysis, 254
- predictive maintenance, 254
- production forecasting, 254
- self-tuning vision systems, 254

industry context, 83

industry-level expertise, 243

insight-to-action arc, 246

integers, types of numbers, 24

interpolations vs extrapolations, 185

interquartile range (IQR), 32, 149–150

inventory holding cost analysis, 256

IQR. *See* interquartile range (IQR)

irregular fluctuations, 151

J

JavaScript Object Notation (JSON) file, 18, 40–41

JOIN, 297

K

- Kaggle, 400
- kicking off an exploratory analysis, 384–385
- KMeans(`n_clusters=4`), 171
- k-means algorithm, 159, 161
- k-means clustering, 160, 254
- k value selection, 162–163

L

- labels, 25–26
- labor utilization analysis, 256
- large language models (LLMs), 11–12, 14, 69, 79, 291, 304
- leadership
 - cases and case studies, 243
 - data-enthusiastic culture, 243–244
 - evangelize company's data projects, 244
 - invest in relationships, 243
 - skills, 239
- lead scoring systems, 268, 280–281
 - behavioral and demographic signals, 281
 - campaign sequence, 281
 - define conversion, 281
 - existing lead database and segment, 281
 - weight assignment, 281
- legal, regulatory, and ethical framework, 352
 - AI ethics framework, 352
 - data privacy laws, 352
 - General Data Protection Regulation (GDPR), 352
- lifetime value (LTV), 267
- linear regression, 144–147, 154–155
 - vs logistic regression, 148, 149
- LinearRegression(), 155, 188–190
 - scikit-learn, 177
- linear regression model
 - constant variance (homoscedasticity), 147
 - independent observations, 146
 - linearity, 146
 - no perfect multicollinearity, 147
 - outliers and leverage points, 147
 - target as numeric, 146
- linear relationship, 30, 105
 - logistic regression, 148
- linear topological structures, 227
- line charts, 224
- live events, 270, 271
- local minimum density, 161
- logic errors, 65–66
- logistical efficiencies, with real-time data, 256

- logistical operations, 256
 - predictive maintenance, 256
 - real-time routing, 256
- logistic regression
 - binary target, 148
 - enough signal and enough events, 148
 - independent observations, 148
 - vs linear regression, 148, 149
 - linear relationship, 148
 - no perfect multicollinearity, 148
 - yes/no decision, 147–149
- LogisticRegression(), 155, 190
- logistics cost-savings demonstration, 258–261
 - AI operational context, 258–259
 - anchor findings, 260–261
 - domain-specific pattern analysis, 259–260
 - push back and pressure-test, 261
- logistics operations, 252
- Looker for Business Intelligence, 194–197
- Looker Studio, 271, 279
- low-stakes decisions, 134
- low-trust scenarios, 71–72

M

- machine learning (ML), 254, 266–268
 - industrial, manufacturing operations
 - defect analysis, 254
 - predictive maintenance, 254
 - production forecasting, 254
 - self-tuning vision systems, 254
 - supervised, 158
 - unsupervised, 158
- machine learning engineers, 365, 366
- machine learning model, 11, 117–120, 124–128, 135–136
 - accuracy metrics, 128–129
 - decision tree, 125–126
 - explaining, 126–128
 - memorizes, 131–132
 - supervised learning, 121
 - classification to predict categories, 122–123
 - comparison with regression and classification, 123–124
 - regression to predict numbers, 121–122
 - training/testing accuracy, 130
 - trust vs escalate decision framework, 134–135
 - underfitting, 132–133
- MAE. *See* Mean Absolute Error (MAE)
- manufacturing operations, 252
 - industrial machine learning
 - defect analysis, 254
 - predictive maintenance, 254

- production forecasting, 254
- self-tuning vision systems, 254
- market basket analysis, 268
- marketing channels, 270, 271
 - AI search, 270
 - live events, 270
 - paid ads, 270
 - referral sites, 270
 - search engine optimization (SEO), 270
 - social media, 270
- marketing-fluent prompting, 274–276
- marketing mix modeling, 268
- marketing optimization, 278
- marketing ROI
 - data and AI products building, 278–279
 - customer segmentation, 279–280
 - lead scoring systems, 280–281
- marketing spend recommendations from ChatGPT, 273–274
- campaign performance data, 274–275
- reallocation, 276–277
- testing AI suggestions, 277–278
- writing fluent prompt, 275–276
- marketing use cases, 266
 - channel scoring, 267
 - content creation, 266–267
 - customer churn reduction, 266
 - dynamic pricing, 267
 - hyper-targeted advertising, 267
 - lead scoring, 268
 - lifetime value forecasting, 267
 - market basket analysis, 268
 - mix modeling, 268
 - recommendation engines, 266
- mass customization production, 254
- Mean Absolute Error (MAE), 146, 181, 182
- mean, statistics, 27–28
- median, statistics, 28
- medium-trust scenarios, 71
- Merriam-Webster Dictionary, 320
- metric specification, 258
- Microsoft Power BI with Copilot, 396
- Microsoft's heavy AI investments, 195
- Missing Completely At Random (MCAR), 31
- Missing Not At Random (MNAR), 31
- missing values, 31, 50–51
- mission statement, 315
- mode, statistics, 28
- multi-tool approach, 64

N

- natural-language querying, 290
 - SQL learning, 297
 - GROUP BY, 297
 - JOIN, 297
 - SELECT, 296–297
 - SQL queries, 291–296
- natural language, querying databases with, 42
- NetSuite, 322
- networking in authentic manner, 369
- neural networks, 60, 125, 254, 343
- nominal categorical data, 25
- nondisclosure agreement (NDA), 243
- non-linear approach, 146
- Not a Number (NaN), 31
- Notion AI for documentation, 398
- novel fraud schemes, 304
- NULL, 31
- number types
 - decimals, 24
 - integers, 24
 - percentages, 24–25

O

- OLS. *See* ordinary least squares (OLS)
- omnichannel analytics, 268–271
- online data visualization design tool, 230
- operational dashboards, 198
- operational improvements, 251
 - AI prompting, 257–261
 - business process automation (BPA), 255
 - logistical operations, 256
 - in logistics, 252
 - in manufacturing, 252
 - operational pain points, 252–253
 - in services, 252
 - in traditional manufacturing, 254–255
- operationalized systems, 279
- operationally fluent prompting, 257–258, 273
 - constraint framing, 258
 - decision anchor, 258
 - domain declaration, 257
 - logistics cost-savings demonstration, 258–261
 - AI operational context, 258–259
 - anchor findings, 260–261
 - domain-specific pattern analysis, 259–260
 - push back and pressure-test, 261
 - metric specification, 258
 - operational pain points, 252–253

- operations datasets, 262
- ordinal categorical data, 25
- ordinary least squares (OLS), 155
- organization
 - business mission, 352
 - business vision, 352
 - company background, 352
 - data science use case, 352
 - operations and activities, 352
 - vision statement, 352
- organizational structure
 - chief executive officer, 352
 - chief information officer, 352
 - data governance committee and officers, 353
 - data privacy compliance program, 353
 - data program/product managers, 353
 - data team roles and responsibilities, 353
- outliers
 - data errors, 150
 - detecting outliers with scipy, 156
 - keep or remove, 150–151
 - leading indicators, 150
 - real but special events, 150
 - types of, 150
 - and unusual data points, 149–151
- outlier value
 - dealing with, 47–48
 - Exploratory Data Analysis (EDA), 104
 - investigating, 52–53
 - signal/error, 32
- overfitting problem, 130–134, 136

P

- packed circle diagrams, 225
- paid ads, 270
- paid ads impressions
 - predicting sales from, 176
 - first-pass regression model, 177–178
 - preparing the data for analysis, 176–177
 - uploading marketing data to ChatGPT ADA, 176
 - vs sales, 183–184
- pairplot(), 115
- pandas, 21, 57, 114–115
 - cleaning data with library, 54–56
- pandas library .corr() method, 154
- pattern recognition, 300–301
 - account takeover fraud, 300
 - application fraud, 301
 - card testing, 300
 - identity fraud, 301
 - velocity fraud, 300–301
- patterns categories, EDA, 101
- Pearson correlation, 154
- people-mapping, 324–325
 - data skillsets, 324–325
 - organizational charts, 324
- percentages, types of numbers, 24–25
- PERCENTILE_CONT, 294
- Perplexity Pro, 399
- personal brand, AI-driven data job, 368
- personally identifiable information, 73
- persuasive data visualizations, 233
- persuasive design, 233
- pie charts, 224–225
- plain-English interpretations, 37
- PLC. *See* Programmable Logic Controller (PLC)
- P&L statement. *See* profit and loss (P&L) statement
- plug-and-play assessments, 338–347
 - AI ethics assessment, 340–345
 - accountability of, 342
 - bias evaluation, 344–345
 - trustworthy AI, 341
 - unbiased results, 342–344
 - data governance assessment, 345–347
 - data privacy policy assessment, 345–347
 - data skills gap analysis, 338–340
- point map, 228, 229
- polynomial regression, 146
- portfolio, AI-driven data job, 367–368
- portfolio pack, 372
- portfolio pack deliverables, 372
 - creating professional-grade visualizations, 373, 375
 - dataset understanding, 374
 - executive summary, 374
 - exploratory analysis, 374
 - segmentation (with metrics), 374
 - visualization, 374
- LinkedIn case study, 376
 - AI-assisted work, 377–378
 - interview walk-through script, preparation, 376–378
- project snapshot, 372
 - approach, 373
 - business problem, 373
 - key findings, 373
 - recommendations, 373
 - technical details, 373
 - showing verification and refinement, 375
 - structuring file and folders, 375–376
- “portfolio-ready” project, 369–370
 - AI tools and prompts, 371
 - business questions, 370
 - business recommendations, 371–372
 - checklist, 369, 370

- complete process, 372
- data sources, 371
- verification, 371
- visualizations, 371
- POTI model, 3
- Power BI, 1, 42, 194–196
 - Copilot, 286, 396
 - integration, 195
 - Desktop, 396
 - Pro, 396
- practical mental models, 60–62
 - confidence, 61
 - generative AI, 60–61
 - tools, 62
- predict categories, classification to, 122–123
- predicting numbers, regression to, 121–122
- predicting sales from paid ads impressions, 176
 - first-pass regression model, 177–178
 - preparing the data for analysis, 176–177
 - uploading marketing data to ChatGPT ADA, 176
- prediction
 - AI uses probability for, 143–144
 - interval, 175
- predictive analytics, 288
 - vs descriptive analysis, 119
 - regression for, 174–176
- predictive model, setting up, 385–386
- prescriptive analytics, 288
- pricing models, 72–73
- proactive AI monitoring, 253
- proactive opportunities, 253
- probability, 142
 - AI uses probability for predictions, 143–144
 - distributions in simple terms, 142–143
- processes, organization, technology, information (POTI) model
 - data science project, 360–362
 - data science use cases, 336–338
- process intelligence, 255
- profitability, 243
- profit and loss (P&L) statement, 244–245
- Programmable Logic Controller (PLC), 254
- progressive disclosure, 200
- prompt engineering, 77
 - adapting, 87–88
- AI
 - analysis formats, 82–83
 - interprets your requests, 79
 - results, 83–84
 - tools, 87–88
- churn analysis, 77, 93, 94, 97
- code, 97–98

- complex questions into simple, 80–82
- context and constraints, 83
- documenting successful prompts for reuse, 86–87
- prompt iteration, 79–80
- prompt templates, 84–85, 91–94
- refinement approach, 95–96
- vague vs specific, 78–79
- workspace in ChatGPT Projects, 88–91
 - analytics, 89
 - approach, 90–91
 - requirements, 90
 - uploading and organizing datasets, 89–90
- prompt iteration, 79–80
- prompt templates, 84–85, 91–94
- psychographics, 217
- push back and pressure-test, 261
- Python code, 1, 3, 12, 14–16, 21, 22, 41, 54, 61, 95, 97, 114, 115, 135
- Python libraries
 - AI for correlation, 154
 - fitting regression with sklearn, 154–155

Q

- quantitative analysis, 24
- querying databases with natural language, 290
 - business intelligence copilots add value, 287
 - generate SQL (structured query language) queries, 291–296
- GROUP BY, 296–297
- JOIN, 296–297
- SELECT, 296–297
- SQL (structured query language), 291
- SQL (structured query language) learning, 296–297
- question core, of template framework, 91
- quick win, 335–338
 - business capability assessment, 335–336
 - processes, organization, technology, information (POTI) model, 336–338
 - selection, 335–336
 - survey, take stock, access, recommend (STAR) framework, 334, 335, 338
- quintessential marketing-fluent prompt, 276

R

- range, measure of spread, 29
- rapid-fire change, 10–11
- raster, 228, 229
- reactive opportunities, 253
- real-time routing, 256
- recommendation engines, 266

- referral sites, 270
- refinement approach, 57, 79–80, 95–96
- regression, 144
 - for business applications, 174
 - comparison with classification and, 123–124
 - fitting regression with sklearn, 154–155
 - linear, 144–147
 - logistic. *See* logistic regression
 - to predicting numbers, 121–122
 - as right tool, 174–176
- regression analysis, 254
 - model's predictions
 - AI to explain R-squared value, 179–180
 - evaluating indicators for trustworthiness, 182
 - “Good fit”, 180–182
 - planning for future scenarios, 184–186
 - for predictive analytics, 174–176
- regression model, 188
 - first-pass, 177–178
- relationships
 - identifying correlations and, 111–113
 - identifying variables and, 104–106
 - using correlation, 30–31
- reliable AI and data science tools, 393–394
 - ChatGPT Advanced Data Analysis, 394
 - Claude by anthropic, 394–395
 - DataCamp and Kaggle for Learning, 399–400
 - Excel with Copilot, 397–398
 - Google BigQuery with Gemini, 397
 - Google Gemini, 395
 - Microsoft Power BI with Copilot, 396
 - Notion AI for documentation, 398
 - perplexity AI for research, 399
- request explanation, of template framework, 91–92
- Return on Ad Spend (ROAS), 274
- revenue distributions, 143
- revenue-generation, 240
- revenue streams vs cost centers, 245
- risk and fraud detection
 - anomaly detection, 301–302
 - definition, 301
 - supervised anomaly detection, 301, 302
 - unsupervised anomaly detection, 301, 302
 - with ChatGPT, 305–309
 - building processes, 308–309
 - flag suspicious activities, 306–307
 - investigating flagged transactions, 307–308
 - transaction data, 306
 - ecommerce and financial services, 300
 - false negative, 302–303
 - false positive, 302–303
 - with generative AI, 303–305
 - with human judgment, 305
 - limitations, 304–305
 - unusual patterns, 303–304
 - overview, 300–303
 - pattern recognition, 300–301
 - account takeover fraud, 300
 - application fraud, 301
 - card testing, 300
 - identity fraud, 301
 - velocity fraud, 300–301
- RMSE. *See* Root Mean Squared Error (RMSE)
- robot workcell, 254
- Root Mean Squared Error (RMSE), 146, 180–182
- R-squared value, 146
 - AI to explain, 179–180

S

- safer workflows, for sensitive data, 74
- sales
 - channels, 267, 268, 270
 - vs paid ads impressions, 183–184
- sanity checks catch errors, 68
- scatter plot, 105, 111, 113, 226–227
- scatter plot matrix, 227
- scikit-learn's LinearRegression(), 177
- scipy, detecting outliers with, 156
- seaborn, 115
- search engine optimization (SEO), 270
- seasonality, time series analysis, 153
- seasonal patterns, 151
- segmentation, analyst, 168–170
- segment names, good and bad, 168
- SELECT, 296–297
- service operations, 252
- Shopify, 266, 305
- silhouette coefficient, 161
- silhouette score, 162–163, 165, 279
- simple questions, 80–82, 107–108
- single-channel predictive model, 274
- skewed distributions, 103, 142
- skills
 - coding, 14, 21–22
 - data visualizations, 218
 - gap analysis, data, 338–340
 - leadership, 239
- skillset and people recommendations
 - data literacy training, 355
 - data team personnel training, 355
 - use case skillset requirements, 355
 - user training, 355

- skillsets, data, 324–325, 336, 338, 339, 358
 - implementation plan, 353
 - inventory, 353
 - methods, 353
- sklearn
 - evaluating models with, 190–191
 - fitting regression with, 154–155
- social media, 270, 271, 279
- spatial plots and maps, 228–229
 - Choropleth, 228, 229
 - point map, 228, 229
 - raster, 228, 229
- specific prompt, vs vague prompt, 78–79
- SQL (structured query language), 291
 - learning, 297
 - GROUP BY, 297
 - JOIN, 297
 - SELECT, 296–297
 - queries generation, 291–296
 - AI to explain, 294–295
 - prompting, 291–294
 - refinement, 296
 - scripts, 1, 3, 42
 - cleaning data, 56–57
 - statements, 292, 294
 - working process, 291
- stacked charts, 225
- stakeholders
 - communicating to, 187–188
 - communication, 215
 - language, 245
 - management, 214–215
 - map, 247
 - prompt structure for, 247–248
- standard chart graphics, 224–225
 - area charts, 224
 - bar charts, 224
 - line charts, 224
 - pie charts, 224–225
- standard deviation, 29, 30, 36
- standard graphics, 225
- StandardScaler(), 171
- STAR framework. *See* survey, take stock, access, recommend (STAR) framework
- statistical interpretation, 387–388
- statistical plots, 226–227
 - histogram, 226
 - scatter plot, 226–227
 - scatter plot matrix, 227
- statistics, 27
 - comparison with mean, median and mode, 27–28
 - intuition, 38
 - spotting relationships using correlation, 30–31
 - spread and variability, 28–30
- step-by-step breakdown, calculation of, 36–37
- strengths, weaknesses, opportunities, and threats (SWOT) analysis, 334, 342
- structured query language (SQL). *See* SQL (structured query language)
- subject matter expertise, business acumen with, 238–239
- subscription models, 72
- summary statement, 351
- Superstore Sales Dashboard, 205
- supervised anomaly detection, 301, 302
- supervised learning, 121
 - classification, 121
 - to predict categories, 122–123
 - comparison with regression and classification, 123–124
 - regression to predict numbers, 121–122
- supervised machine learning, 158
- surveys and interviews
 - business managers, 326
 - data professionals, 326
 - data resources, 321–322
 - data technologies, 318, 319
 - decision-makers, 326
 - employees and contractors, 326
 - human resource managers, 326
 - IT professionals, 326
 - potential stakeholders, 326
 - sales and marketing professionals, 326
- survey, take stock, access, recommend (STAR) framework, 316–317, 334, 335, 338
- synthetic dataset, 306
- systematic exploration, with Claude, 108–114

T

- Tableau, 1, 194–196
 - Creator, 396
 - Einstein Copilot, 196
 - Public, 397
 - with Tableau AI, 396–397
- target audience, designing for, 216
 - brainstorming, 216–218
 - choose functional design, 216, 218–219
 - visualization, 216, 218
- testing accuracy, 130, 132, 133
- text cleaning, 45–47
- text data, 27
- three-sigma rule, 32
- threshold tuning, 303
- time-based analysis, 26

- time format
 - context, 83
 - dealing with, 26
 - standardizing, 44–45
- time series analysis, 151
 - seasonality and forecasting, 153
 - trends and patterns over time, 151–153
- time series decomposition, 153
- tools, AI, 1, 3–5, 257
 - comparison with ChatGPT, Claude, and Gemini, 62–63
 - data categories, 73–74
 - data minimization, 73
 - failure modes for, 65–67
 - assumption errors, 66
 - catching calculation and logic errors, 65–66
 - data-type and unit mistakes, 66–67
 - hallucination patterns, 65
 - workflow failures, 67
 - fast verification techniques, 67–70
 - multi-tool approach, 64
 - practical mental models, 60–62
 - pricing models, 72–73
 - right tool for different tasks, 63–64
 - safer workflows, 74
 - trust rubric, 70–72
- topology structures, 227–228
 - graph models, 227
 - linear topological structures, 227
 - tree network topology, 227
- traditional business intelligence (BI), 285
- traditional manufacturing, operational improvements in, 254–255
- training accuracy, 130–133
- treemaps, 225
- tree network topology, 227
- trend categories, EDA, 101
- trust rubric, 70–72, 305
 - high-trust scenarios, 71
 - human expertise, 72
 - low-trust scenarios, 71–72
 - medium-trust scenarios, 71
- trust vs escalate decision framework, 134–135

U

- underfitting, 132–134
- unsupervised anomaly detection, 301, 302
- unsupervised learning, 121

- unsupervised machine learning, 158
- unusual data points, outliers and, 149–151
- usage-based models, 72
- use cases in marketing, 266
 - channel scoring, 267
 - content creation, 266–267
 - customer churn reduction, 266
 - dynamic pricing, 267
 - hyper-targeted advertising, 267
 - lead scoring, 268
 - lifetime value forecasting, 267
 - market basket analysis, 268
 - mix modeling, 268
 - recommendation engines, 266

V

- vague prompt, vs specific prompt, 78–79
- `.value_counts()`, 114–115
- variance, of standard deviation, 29
- velocity fraud, 300–301
- verification literacy, 59, 74
- vibe-coding a dashboard in lovable, 201–202
 - building a dashboard, 203–209
 - examining the dataset, 202–203
- visualization, 219. *See also* data visualizations
 - designing, 386–387
 - for target audience, 216, 218
- VP of Finance, 247, 248
- VP of Operations, 241, 245, 251
- VP of Product, 247, 248

W

- waste reduction and resource allocation, 256–257
- web analytics into dollars and sense, 268–269
 - channel mapping, 269–270
 - channel performance, 270–271
 - channel scoring, 271–273
 - omnichannel analytics, 269
- whole numbers (integers), 24
- word clouds, 225
- workflow failures, 67
- workflow mining, 255

Z

- zigzag model, 131