# Week 5 (Tue):

## e-Business

## Semester 2, 2025

- **Today's Goal**
    - **Basic Framework to Understand Data Science**
    - **Google Colab Practice I**

# Business Analytics

- The Converging Forces: Platform Systems + Smartphones
  - Digital platforms (e.g., Naver Smart Store, Coupang, App Store, Google Play)
  - Ubiquitous smartphones as always-on access points

- Resulting Shift
  - Commerce, communication, and service delivery now occur inside integrated platform–app ecosystems
  - Customer interactions are continuous, multi-channel, and data-rich

- Key functions inside platform-based apps:
  - CRM and direct messaging
  - Loyalty & membership management
  - Seamless ordering, payment, and subscription

- Distribution Disruption
  - Firms reach customers directly through owned apps and platform storefronts
  - Platforms lower entry barriers so even small firms can operate direct-to-consumer

- Access is no longer scarce; relationship quality and retention matter most

# The Basic Framework

- Why Business Analytics is Mission-Critical

  - Data Explosion

  - Platform and app interactions generate high-volume, high-velocity data

  - Analytical Needs

  - Map and predict customer journeys across devices and touchpoints

  - Measure network effects and cross-side platform dynamics

  - Personalize offers and content in real time

*Performance = **f**( Business, IT )*

# Common Business Analytics Problems

- How likely is client X to buy product Y?
- Which clients are “at risk” of going to our competitors?
- What kind of promotions should we offer to retain our customers?

  → These questions commonly ask the relationship between X and Y for business analytics.

- Examples:
  - CIO needs to decide whether to purchase AI systems.
  - HR needs to decide whether to promote you.
  - HR wants to decide how much bonus to increase.
  - When a user comes to Amazon, Amazon needs to decide which products to recommend.

# Business Case

The drive to capture and organize customer information is longstanding, but modern platform retailers such as Amazon have turned it into a sophisticated, analytics-driven process. For decades, Amazon has built extensive data reservoirs on every shopper who regularly interacts with its physical or digital storefronts. Whenever possible, the company assigns each customer a persistent identifier—internally called a “guest ID”—that becomes the central key for integrating data from many touchpoints.

A manager explains:

“If you use a credit card or coupon, complete a survey, request a refund, contact our help line, open one of our emails, or visit our website, each event is captured and tied back to your guest ID. Our goal is to understand every dimension of the customer journey.”

# Business Case

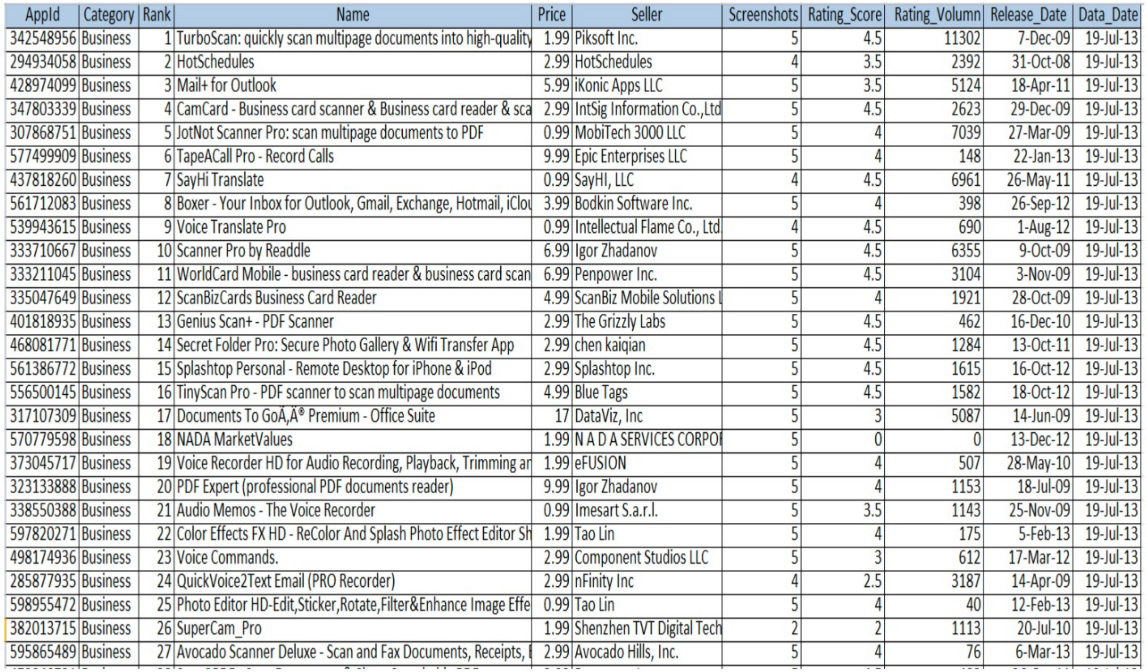
This ID then serves as the anchor for a rich, constructed dataset that blends first-party and third-party sources. Linked attributes include basic demographics (age, marital and family status, neighborhood, commute time, income estimates) and financial indicators (credit card types, recent moves, credit risk flags such as bankruptcy or divorce). Amazon supplements these with purchased or inferred data—educational background, home purchase year, preferred brands, online discussion topics, magazine subscriptions, charitable giving, political leanings, and even the number of cars owned.

Through this process, business analytics does more than store transactions: it constructs a multi-layered customer graph, enabling predictive modeling for personalized recommendations, dynamic pricing, targeted promotions, and long-term relationship management. In platform ecosystems where mobile apps and web interactions create continuous data flows, such data construction, is the foundation on which competitive analytics and customer insight are built.

# Business Case

- Record data: a collection of records, each of which consists of a fixed set of attributes
- Transaction data: Each record(transaction) involves a set of items

Variables (or Attributes / Features)

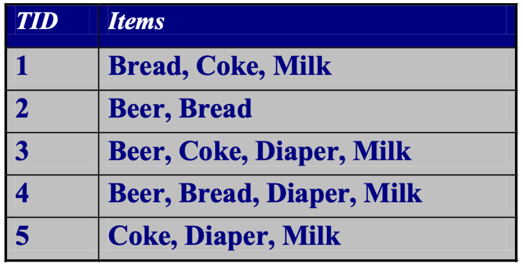


| AppId | Category | Rank | Name | Price | Seller | Screenshots | Rating_Score | Rating_Volumn | Release_Date | Data_Date |
|---|---|---|---|---|---|---|---|---|---|---|
| 342548956 | Business | 1 | TurboScan: quickly scan multipage documents into high-quality | 1.99 | Piksoft Inc. | 5 | 4.5 | 11302 | 7-Dec-09 | 19-Jul-13 |
| 294934058 | Business | 2 | HotSchedules | 2.99 | HotSchedules | 4 | 3.5 | 2392 | 31-Oct-08 | 19-Jul-13 |
| 428974099 | Business | 3 | Mail+ for Outlook | 5.99 | iKonic Apps LLC | 5 | 3.5 | 5124 | 18-Apr-11 | 19-Jul-13 |
| 347803339 | Business | 4 | CamCard - Business card scanner & Business card reader & sca | 2.99 | IntSig Information Co.,Ltd | 5 | 4.5 | 2623 | 29-Dec-09 | 19-Jul-13 |
| 307868751 | Business | 5 | JotNot Scanner Pro: scan multipage documents to PDF | 0.99 | MobiTech 3000 LLC | 5 | 4 | 7039 | 27-Mar-09 | 19-Jul-13 |
| 577499909 | Business | 6 | TapeACall Pro - Record Calls | 9.99 | Epic Enterprises LLC | 5 | 4 | 148 | 22-Jan-13 | 19-Jul-13 |
| 437818260 | Business | 7 | SayHi Translate | 0.99 | SayHI, LLC | 4 | 4.5 | 6961 | 26-May-11 | 19-Jul-13 |
| 561712083 | Business | 8 | Boxer - Your Inbox for Outlook, Gmail, Exchange, Hotmail, iClou | 3.99 | Bodkin Software Inc. | 5 | 4 | 398 | 26-Sep-12 | 19-Jul-13 |
| 539943615 | Business | 9 | Voice Translate Pro | 0.99 | Intellectual Flame Co., Ltd | 4 | 4.5 | 690 | 1-Aug-12 | 19-Jul-13 |
| 333710667 | Business | 10 | Scanner Pro by Readdle | 6.99 | Igor Zhadanov | 5 | 4.5 | 6355 | 9-Oct-09 | 19-Jul-13 |
| 333211045 | Business | 11 | WorldCard Mobile - business card reader & business card scan | 6.99 | Penpower Inc. | 5 | 4.5 | 3104 | 3-Nov-09 | 19-Jul-13 |
| 335047649 | Business | 12 | ScanBizCards Business Card Reader | 4.99 | ScanBiz Mobile Solutions L | 5 | 4 | 1921 | 28-Oct-09 | 19-Jul-13 |
| 401818935 | Business | 13 | Genius Scan+ - PDF Scanner | 2.99 | The Grizzly Labs | 5 | 4.5 | 462 | 16-Dec-10 | 19-Jul-13 |
| 468081771 | Business | 14 | Secret Folder Pro: Secure Photo Gallery & Wifi Transfer App | 2.99 | chen kaiqian | 5 | 4.5 | 1284 | 13-Oct-11 | 19-Jul-13 |
| 561386772 | Business | 15 | Splashtop Personal - Remote Desktop for iPhone & iPod | 2.99 | Splashtop Inc. | 5 | 4.5 | 1615 | 16-Oct-12 | 19-Jul-13 |
| 556500145 | Business | 16 | TinyScan Pro - PDF scanner to scan multipage documents | 4.99 | Blue Tags | 5 | 4.5 | 1582 | 18-Oct-12 | 19-Jul-13 |
| 317107309 | Business | 17 | Documents To GoÃ‚Â® Premium - Office Suite | 17 | DataViz, Inc | 5 | 3 | 5087 | 14-Jun-09 | 19-Jul-13 |
| 570779598 | Business | 18 | NADA MarketValues | 1.99 | N A D A SERVICES CORPOR | 5 | 0 | 0 | 13-Dec-12 | 19-Jul-13 |
| 373045717 | Business | 19 | Voice Recorder HD for Audio Recording, Playback, Trimming ar | 1.99 | eFUSION | 5 | 4 | 507 | 28-May-10 | 19-Jul-13 |
| 323133888 | Business | 20 | PDF Expert (professional PDF documents reader) | 9.99 | Igor Zhadanov | 5 | 4 | 1153 | 18-Jul-09 | 19-Jul-13 |
| 338550388 | Business | 21 | Audio Memos - The Voice Recorder | 0.99 | Imesart S.a.r.l. | 5 | 3.5 | 1143 | 25-Nov-09 | 19-Jul-13 |
| 597820271 | Business | 22 | Color Effects FX HD - ReColor And Splash Photo Effect Editor Sh | 1.99 | Tao Lin | 5 | 4 | 175 | 5-Feb-13 | 19-Jul-13 |
| 498174936 | Business | 23 | Voice Commands. | 2.99 | Component Studios LLC | 5 | 3 | 612 | 17-Mar-12 | 19-Jul-13 |
| 285877935 | Business | 24 | QuickVoice2Text Email (PRO Recorder) | 2.99 | nFinity Inc | 4 | 2.5 | 3187 | 14-Apr-09 | 19-Jul-13 |
| 598955472 | Business | 25 | Photo Editor HD-Edit,Sticker,Rotate,Filter&Enhance Image Effe | 0.99 | Tao Lin | 5 | 4 | 40 | 12-Feb-13 | 19-Jul-13 |
| 382013715 | Business | 26 | SuperCam_Pro | 1.99 | Shenzhen TVT Digital Tech | 2 | 2 | 1113 | 20-Jul-10 | 19-Jul-13 |
| 595865489 | Business | 27 | Avocado Scanner Deluxe - Scan and Fax Documents, Receipts, I | 2.99 | Avocado Hills, Inc. | 5 | 4 | 76 | 6-Mar-13 | 19-Jul-13 |

ID

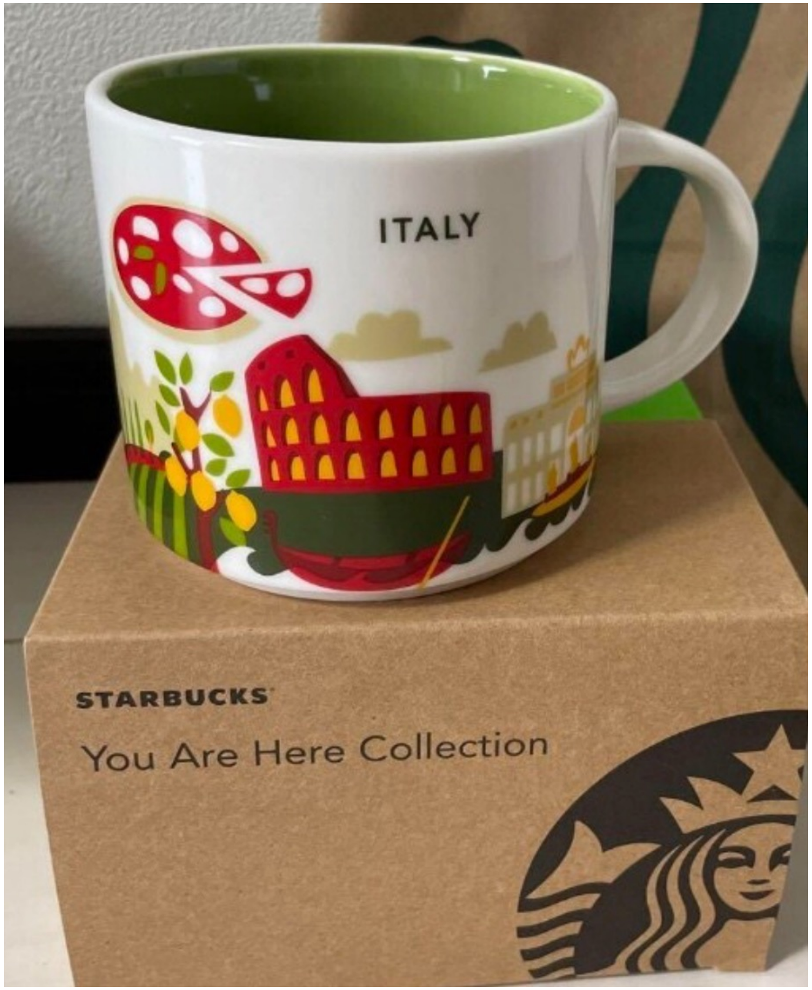
Categorical Variables

Numeric Variables



| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

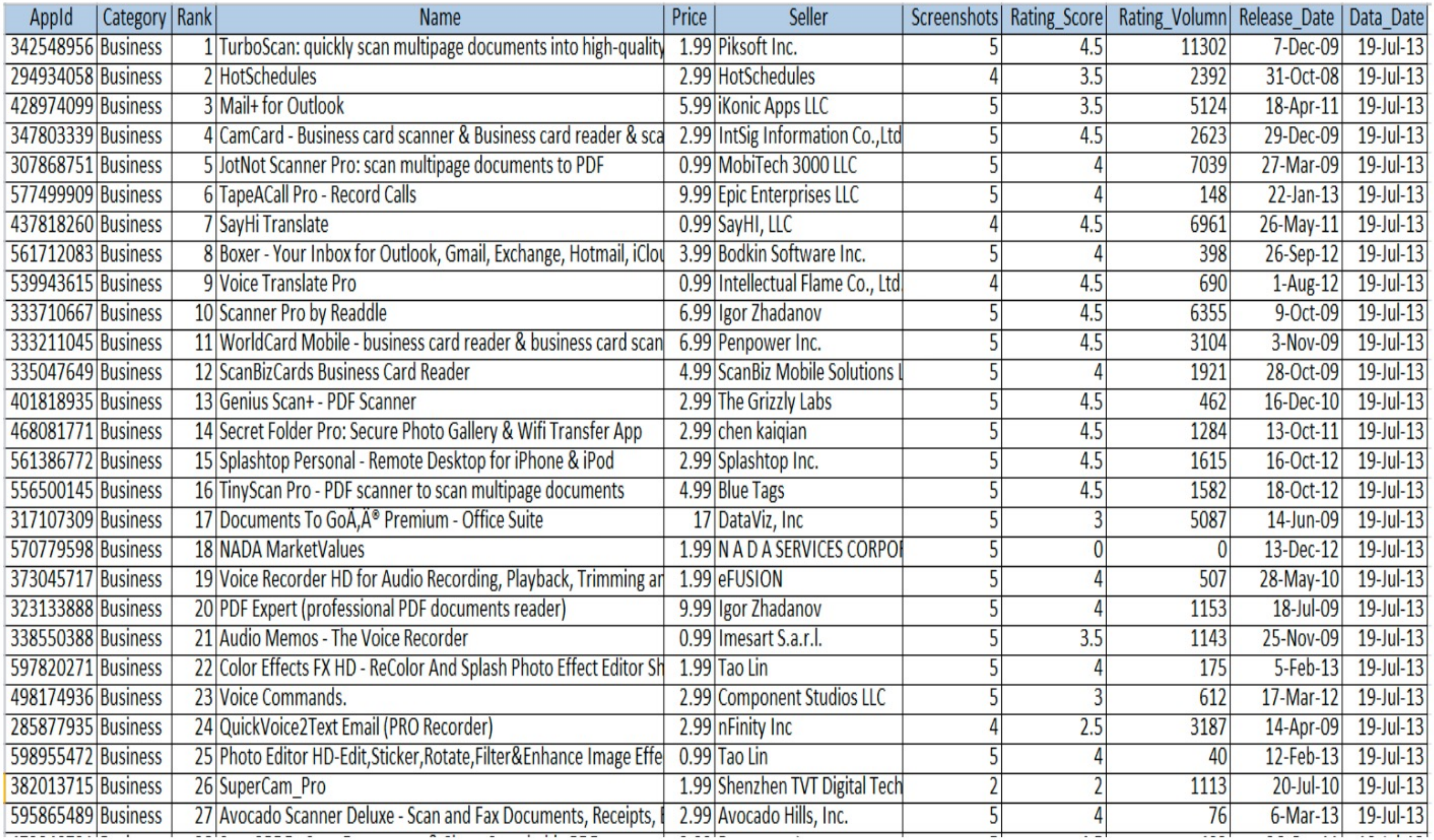
# Describe the Present to Your Parents or Friends



# Descriptive Analytics

- Identify, classify, and count objects or events
  - Money spent or number of widgets manufactured in the past time period.
  - They are critical for knowing how the organization is performing—your current situation.
  - For example, they let you know if you are above or below budget, or are performing up to standards.
  - This category includes reports and dashboards.

# What Can You Expect?



| AppId | Category | Rank | Name | Price | Seller | Screenshots | Rating_Score | Rating_Volumn | Release_Date | Data_Date |
|---|---|---|---|---|---|---|---|---|---|---|
| 342548956 | Business | 1 | TurboScan: quickly scan multipage documents into high-quality | 1.99 | Piksoft Inc. | 5 | 4.5 | 11302 | 7-Dec-09 | 19-Jul-13 |
| 294934058 | Business | 2 | HotSchedules | 2.99 | HotSchedules | 4 | 3.5 | 2392 | 31-Oct-08 | 19-Jul-13 |
| 428974099 | Business | 3 | Mail+ for Outlook | 5.99 | iKonic Apps LLC | 5 | 3.5 | 5124 | 18-Apr-11 | 19-Jul-13 |
| 347803339 | Business | 4 | CamCard - Business card scanner & Business card reader & sca | 2.99 | IntSig Information Co.,Ltd | 5 | 4.5 | 2623 | 29-Dec-09 | 19-Jul-13 |
| 307868751 | Business | 5 | JotNot Scanner Pro: scan multipage documents to PDF | 0.99 | MobiTech 3000 LLC | 5 | 4 | 7039 | 27-Mar-09 | 19-Jul-13 |
| 577499909 | Business | 6 | TapeACall Pro - Record Calls | 9.99 | Epic Enterprises LLC | 5 | 4 | 148 | 22-Jan-13 | 19-Jul-13 |
| 437818260 | Business | 7 | SayHi Translate | 0.99 | SayHI, LLC | 4 | 4.5 | 6961 | 26-May-11 | 19-Jul-13 |
| 561712083 | Business | 8 | Boxer - Your Inbox for Outlook, Gmail, Exchange, Hotmail, iClou | 3.99 | Bodkin Software Inc. | 5 | 4 | 398 | 26-Sep-12 | 19-Jul-13 |
| 539943615 | Business | 9 | Voice Translate Pro | 0.99 | Intellectual Flame Co., Ltd | 4 | 4.5 | 690 | 1-Aug-12 | 19-Jul-13 |
| 333710667 | Business | 10 | Scanner Pro by Readdle | 6.99 | Igor Zhadanov | 5 | 4.5 | 6355 | 9-Oct-09 | 19-Jul-13 |
| 333211045 | Business | 11 | WorldCard Mobile - business card reader & business card scan | 6.99 | Penpower Inc. | 5 | 4.5 | 3104 | 3-Nov-09 | 19-Jul-13 |
| 335047649 | Business | 12 | ScanBizCards Business Card Reader | 4.99 | ScanBiz Mobile Solutions L | 5 | 4 | 1921 | 28-Oct-09 | 19-Jul-13 |
| 401818935 | Business | 13 | Genius Scan+ - PDF Scanner | 2.99 | The Grizzly Labs | 5 | 4.5 | 462 | 16-Dec-10 | 19-Jul-13 |
| 468081771 | Business | 14 | Secret Folder Pro: Secure Photo Gallery & Wifi Transfer App | 2.99 | chen kaiqian | 5 | 4.5 | 1284 | 13-Oct-11 | 19-Jul-13 |
| 561386772 | Business | 15 | Splashtop Personal - Remote Desktop for iPhone & iPod | 2.99 | Splashtop Inc. | 5 | 4.5 | 1615 | 16-Oct-12 | 19-Jul-13 |
| 556500145 | Business | 16 | TinyScan Pro - PDF scanner to scan multipage documents | 4.99 | Blue Tags | 5 | 4.5 | 1582 | 18-Oct-12 | 19-Jul-13 |
| 317107309 | Business | 17 | Documents To GoĂ,Ă® Premium - Office Suite | 17 | DataViz, Inc | 5 | 3 | 5087 | 14-Jun-09 | 19-Jul-13 |
| 570779598 | Business | 18 | NADA MarketValues | 1.99 | N A D A SERVICES CORPOI | 5 | 0 | 0 | 13-Dec-12 | 19-Jul-13 |
| 373045717 | Business | 19 | Voice Recorder HD for Audio Recording, Playback, Trimming ar | 1.99 | eFUSION | 5 | 4 | 507 | 28-May-10 | 19-Jul-13 |
| 323133888 | Business | 20 | PDF Expert (professional PDF documents reader) | 9.99 | Igor Zhadanov | 5 | 4 | 1153 | 18-Jul-09 | 19-Jul-13 |
| 338550388 | Business | 21 | Audio Memos - The Voice Recorder | 0.99 | Imesart S.a.r.l. | 5 | 3.5 | 1143 | 25-Nov-09 | 19-Jul-13 |
| 597820271 | Business | 22 | Color Effects FX HD - ReColor And Splash Photo Effect Editor Sh | 1.99 | Tao Lin | 5 | 4 | 175 | 5-Feb-13 | 19-Jul-13 |
| 498174936 | Business | 23 | Voice Commands. | 2.99 | Component Studios LLC | 5 | 3 | 612 | 17-Mar-12 | 19-Jul-13 |
| 285877935 | Business | 24 | QuickVoice2Text Email (PRO Recorder) | 2.99 | nFinity Inc | 4 | 2.5 | 3187 | 14-Apr-09 | 19-Jul-13 |
| 598955472 | Business | 25 | Photo Editor HD-Edit,Sticker,Rotate,Filter&Enhance Image Effe | 0.99 | Tao Lin | 5 | 4 | 40 | 12-Feb-13 | 19-Jul-13 |
| 382013715 | Business | 26 | SuperCam_Pro | 1.99 | Shenzhen TVT Digital Tech | 2 | 2 | 1113 | 20-Jul-10 | 19-Jul-13 |
| 595865489 | Business | 27 | Avocado Scanner Deluxe - Scan and Fax Documents, Receipts, I | 2.99 | Avocado Hills, Inc. | 5 | 4 | 76 | 6-Mar-13 | 19-Jul-13 |

# Insights for Common BA Problems

- How likely is client X to buy product Y?
  - After knowing client X’s likelihood of purchasing the product Y, what a firm can do next?
- Which clients are “at risk” of going to our competitors?
  - After knowing clients’ propensity to leave, what is your suggestion to the firm?
- What kind of promotions should we offer to retain our customers?
  - After knowing effectives of promotions, what are the implications?

# Predictive Analytics

- Look at the trend of past events to anticipate possible future outcomes.
  - This allows the organization to better plan for the future—deciding what actions to take that can improve the future results.
  - Discover patterns and correlations in data that might be missed by the human eye.
  - Build a model that will allow “What if” analyses to decide on the best course of action.

# Values of Business Analytics

- **Identify hidden patterns & unknown correlations**

  - Extract insights from massive, heterogeneous data sets
  - Reveal relationships not visible to human intuition

- Enable better business decisions

  - Guide strategic planning and operational execution
  - Support forecasting and risk management
  - Drive financial and operational performance

- But BA is hard

  - Data originate from multiple, often incompatible sources
  i) internal systems (transactions, CRM, app logs)
  ii) external feeds (social media, sensors, third-party data)

  - Integrating, cleaning, and governing these streams is complex
  - Hidden patterns may be deep, nonlinear, and rare, requiring advanced analytics and continual model refinement

# Features of Business Analytics

- Sometime, analysts may throw out variables that seem unlikely to be interesting, keeping only a few carefully chosen variable they expect to be important.
- The business analytic approach calls for letting the data itself reveal what is and is not important.
- At the beginning, we should use many variables, and let the model decide.
- It depends on the particular algorithms employed, the complexity of data.
- When data is scares, business analytics is less effective and it is less likely to be useful.
- Then, where should we start?

# Basics of Business Analytics

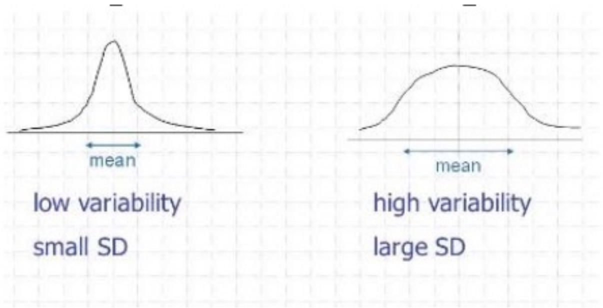
- The mean is the most common measure of the location of a set of points
- However, the mean is very sensitive to outliers
- Thus, the median is also commonly used.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

# Basics of Business Analytics

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points
- Standard deviation is the spread of a group of numbers from the mean
- Variance measures the average degree to which each point differs from the mean
- While standard deviation is the square root of the variance, variance is the average of all data points within a group

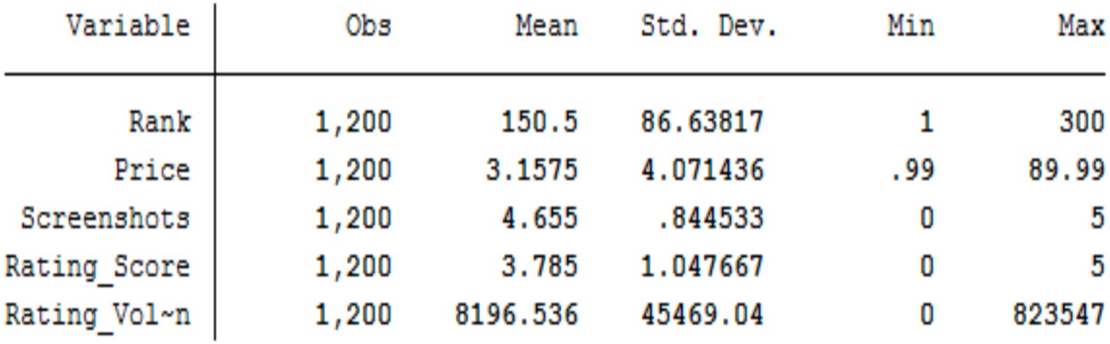


**Variance** $$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

**Standard deviation** $$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

# Basics of Business Analytics

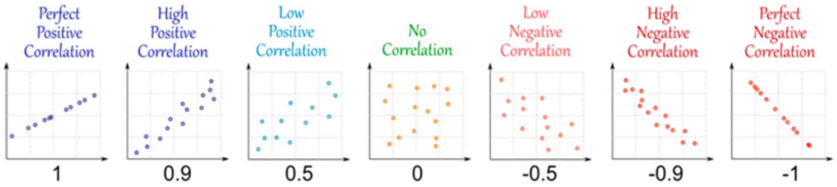
- Report format in general for summary statistics or descriptive statistics

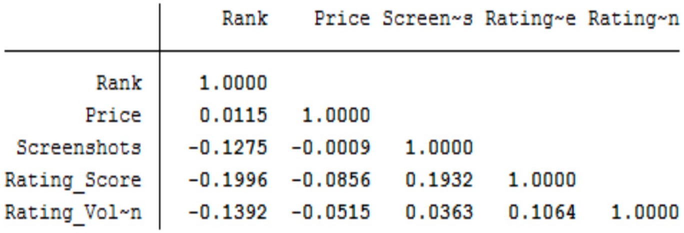


```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        Rank |      1,200       150.5    86.63817          1        300
       Price |      1,200      3.1575    4.071436        .99      89.99
 Screenshots |      1,200       4.655     .844533          0          5
Rating_Score |      1,200       3.785    1.047667          0          5
Rating_Vol~n |      1,200    8196.536    45469.04          0     823547
```

# Basics of Business Analytics

- In statistics, dependence is any statistical relationship between two variables or two sets of data
- Correlation refers to any of a broad class of statistical relationships involving dependence.
- The most familiar measure of dependence between two quantities is Pearson's correlation coefficient.
- It is obtained by dividing the covariance of the two variables by the product of their standard deviations.





```
             |     Rank    Price Screen~s Rating~e Rating~n
-------------+---------------------------------------------
        Rank |   1.0000
       Price |   0.0115   1.0000
 Screenshots |  -0.1275  -0.0009   1.0000
Rating_Score |  -0.1996  -0.0856   0.1932   1.0000
Rating_Vol~n |  -0.1392  -0.0515   0.0363   0.1064   1.0000
```

# Basics of Business Analytics

- Data quality problems that affect the mean and deviations

  - Duplicate data

  - Outliers

  - Missing values

- What can we do about these problems?

# Basics of Business Analytics

- Duplicate data: The data set may include data objects that are duplicates

  - Major issue when merging data from heterogeneous sources

- Example: Same person with multiple email addresses
- Data cleaning method

  - Merge some

  - Delete

  - Keep the newest one

# Basics of Business Analytics

- Missing values

  - When information is not collected (e.g., people decline to give their age and weight)

  - Attributes may not be applicable to all cases.

- Special case of “0”

  - The value is missing

  - The value is actually zero.

- Most algorithms will not process records with missing values.

# Basics of Business Analytics

- Eliminate data records

  - If a small number of records have missing values, you can omit them.

- Imputation

  - 1 variable with many missing values, and 29 variables without missing values

  - Replace missing values with reasonable substitutes (e.g., mean)

  - Let’s keep the record and use the rest of its information(e.g., non-missing).

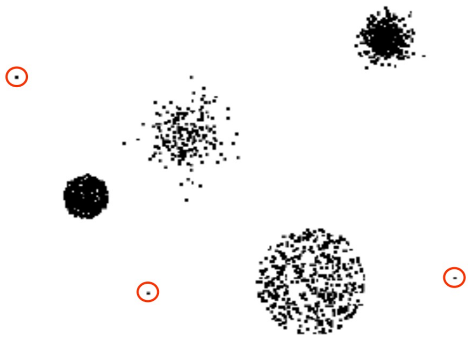
- Eliminate data variables

  - If many records are missing values on a small set of variables, you can drop those variables. If many records have missing values, omission is not practical.

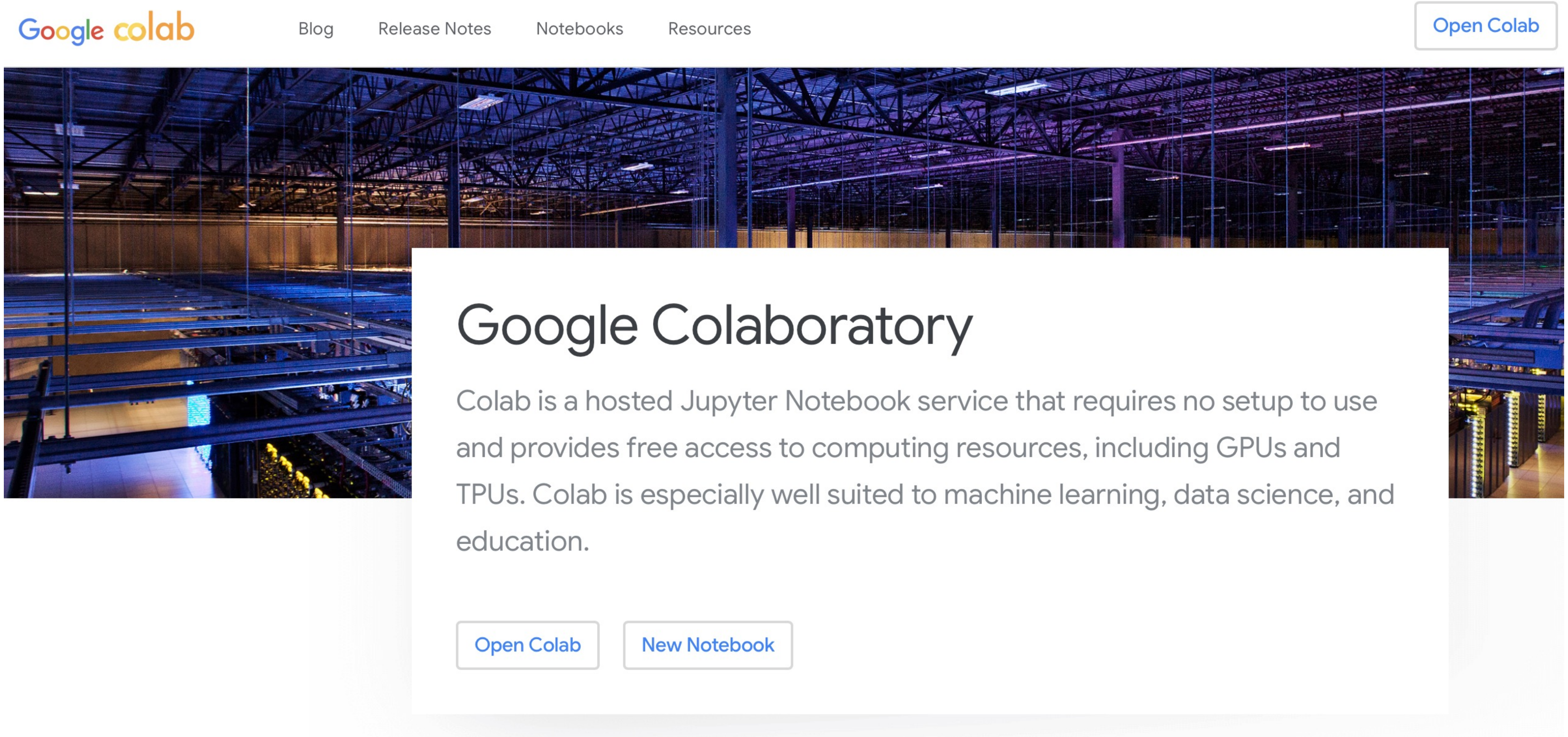
# Basics of Business Analytics

- Outliers are data objects with characteristics that are considerably different than most of the other data objects.
- An outlier is an observation that is extreme, being distant from the rest of the data.
- Outliers can have disproportionate influence on models.
- An important step in data pre-processing is detecting outliers.
- Once detected, domain knowledge is required to determine if it is an error, or truly extreme.

  - For example, temperature of 40 Celsius degree in Korea.



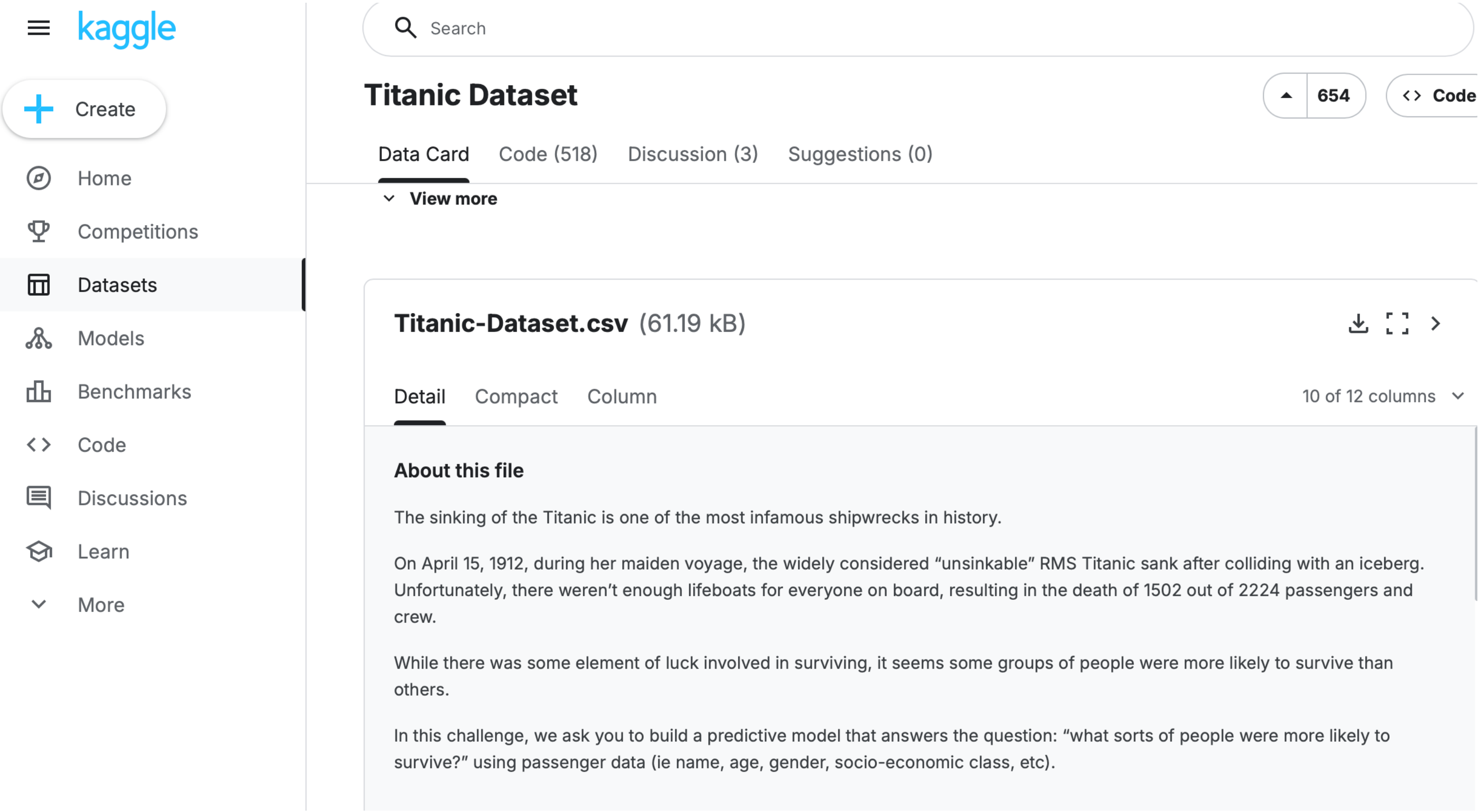
# Practice



## Google Colaboratory

Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education.

# Practice



## Titanic Dataset

**Titanic-Dataset.csv** (61.19 kB)

**About this file**

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (ie name, age, gender, socio-economic class, etc).

# Homework I (Due Date: 1pm Oct 14)

- Deliverables - A single Jupyter Notebook (.ipynb) with:

  - Name and ID

  - All code cells executed

  - Plots displayed

  - Try to answer to all questions in markdown cells

  (If you are unsure of an answer, leave the section blank.)

- Questions

  - Question 1: How many rows and columns does the dataset have?

  - Question 2: Which variables are numeric? Which are categorical?

  - Question 3: What is the mean and median of Age and Fare?

  - Question 4: Which passenger class (Pclass) had the most travelers?

  - Question 5: How did removing outliers affect the shape of Fare's distribution?

  - Question 6: Describe the shape of the age distribution.

  - Question 7: How do survival rates differ by passenger class?