Introductions

Instructor: Kasun Samarsinghe
Sr. Data Scientist, Device Authentication and Identification, American Express
Background: Security in Fintech and e-Commerce
Hobbies: Biking and Hiking

Introduce yourself!!
Your Department and year/ quarter
Past data related experience – aspirations
Hobbies

Group Work

- Assignments should be completed as group work and each group should have at least 4 students each
  - *Exceptions will be given for 3 student groups predicated on the class size*
- Please make sure that each individual group member is contributing towards assignments and the final project
- Once a group is formed, one member of the group should email me with group member names;
  - Each group should submit 1 report (with the code) for assignment work
  - Every report should contain all group member names
- Group work will be assessed during the 4 Assignments and the Final Project
  - Data selected for Assignment1 can be used for later Assignments and the Final Project
  - Choose a dataset with high cardinality, variability, diversity of data types, larger size etc…
- Week 5 will be an in-class Quiz – will be individually assessed
- Discussion: Initial posting, end of Day 3 (11:59 PM, ET). Secondary posting, to at least 2 of your classmate's postings by the end of Day 7 (11:59 PM, ET).

How do we get started

- Collection of data can be explained in two main methods
  - Descriptive statistics – Take a sample/ group and record statistics such as mean, median etc. for that group
  - Inferential statistics – Select a representative sample from a population and derive inferences about the larger population’s statistics

- Where can we find the data:
  - Google dataset search: https://datasetsearch.research.google.com/
  - data.world - https://data.world/
  - Github/ Kaggle -
  - Scarping websites and via API
  - Industry
    - Database: SQL, NoSQL, Oracle, MongoDB, Cassandra
    - API calls – time consuming, resource extensive
    - Colleague: local files (using SFTP)
    - Web scraping: running a JS over the website to scrape web data (Eg: urllib and beautifulsoup)
    - AWS: S3 buckets, DynamoDB, SimpleDB

What is the structure of your data?

- How Much Of A Mess Is Your Data?
  - Is The Data Structured?
    - Structured data: Database – Primary Key, Secondary and Composite Keys are pre-defined
    - CSV – schema is pre-defined
  - Unstructured Data:
    - JSON - schema can change over the data points
    - XML
    - Parquet
    - What is the difference between JSON and Parquet file formats?
      - JSON is easier to write, but takes a longer time to read compared to Parquet
    - Text
      - Should be parsed using pre-defined dictionaries to make sense out of the data

Data Cleanup and Pre-Processing

- Data Scientists, typically spent 50% of their time on cleaning and understanding the data
  - Important to understand the data dictionary and the problem domain
- What Do We Need To Look For
  - Missing Values
  - Anomalies
  - Typos
  - Class Imbalance – Fixed by Over sampling, Under sampling, Synthetic sampling (SMOTE)
- Type Conversion
  - Variables can be of type: Numerical, Categorical, Ordinal, Date, Character etc.
  - 5 – Could be a numerical, categorical or ordinal variable – convert to the correct type based on the data
  - 2018-01-05 – Convert this to a date type
  - R has as.factor, as.numeric, as.Date functions that help type conversions

??What is the difference between Data Engineers and Data Scientists

Missing Values: How to handle them?

- Most real-life data is not complete
- The missing data should be handled in such a way that it doesn't affect the credibility of the data and the model concept
- Most tree based models can handle missing data
- Techniques:
  - If the missing value rows are ~0-0.5% - remove the examples
  - If the missing values rows are ~0.5-20% - replace with mean (normally distributed)/ median(imbalanced data)
  - Encode the missing values to be -1 for tree based models
  - Impute missing values using predictive models or clustering – k Nearest Neighbors Imputer
  - Replace NA's with 0

Anomalies (Outliers/ Novelties)

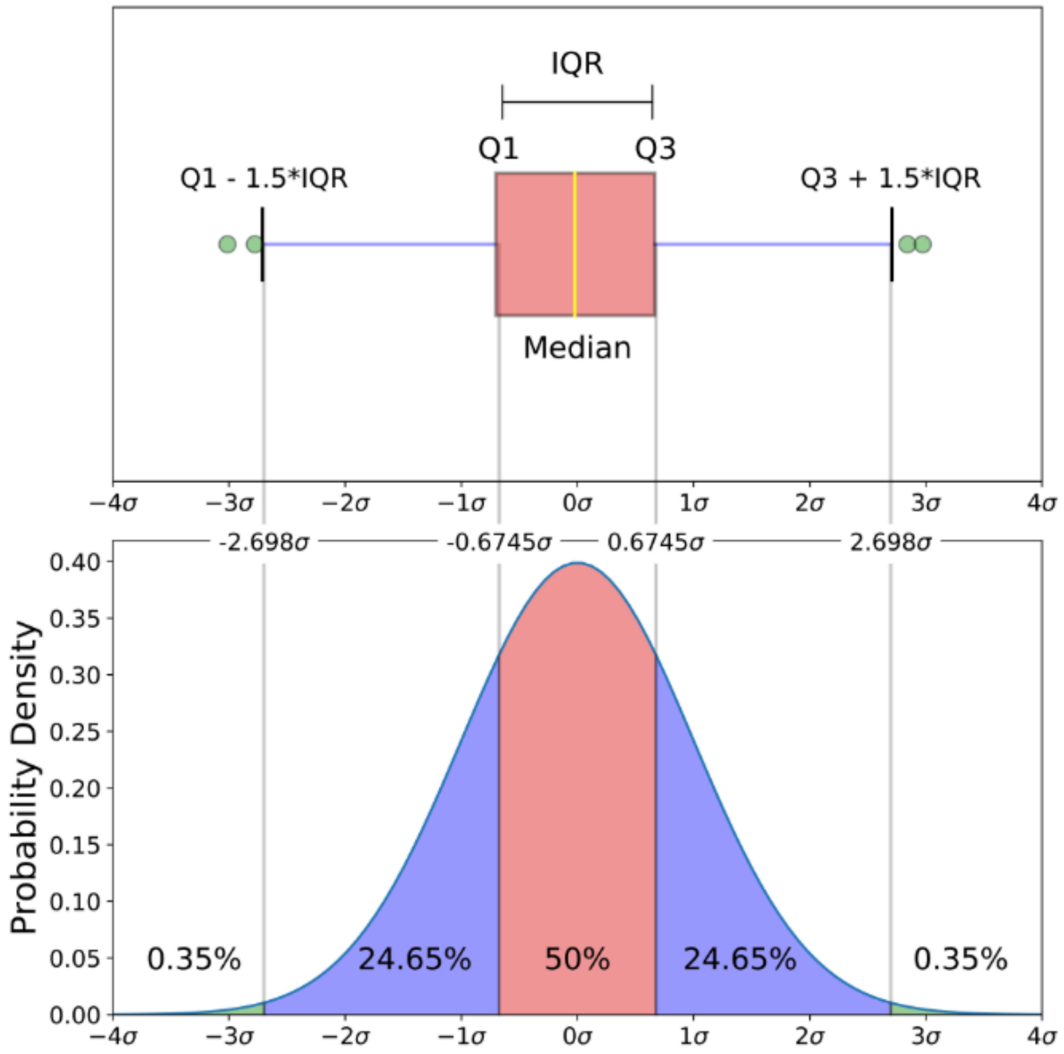
- Observations that are outside of the general distribution of the normal observations
  - Outlier detection: Detecting deviant observations when training data is polluted with outlier observations – Eg: Multivariate Gaussian Distribution
  - Novelty detection: Training data is not polluted and detecting whether a NEW observation is an outlier or not – Eg: 1-class SVM
- Extraneous values in a numerical variable can be detected using
  - Observations that are outside of certain factors of the std. deviation
  - Boxplot
- Text based anomalies
  - Misspellings, case sensitives, white spaces, date conversions etc.
- Before modeling, outliers should be removed from the dataset

BoxPlot

- **IQR:** Inter quartile range: Q3 – Q1

- **“maximum”**: Q3 + 1.5*IQR
- **“minimum”**: Q3 - 1.5*IQR

- Observe that the outliers are all the observations that are outside of the **maximum** and **minimum** points



Comparison of a boxplot of a nearly normal distribution and a probability density function (pdf) for a normal distribution

Source: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Exploratory Data Analysis: EDA

- Processed cleaned data can now be analyzed
  - Summary statistics: Mean, Median, Mode, Frequency Distribution
  - Correlation matrix – relationship between variables
  - Use summary(data) function in R
  - Central Limit Theorem
- Visualize your findings
  - Single Variable
    - Histograms/ Frequency plots
    - Boxplots/ Outlier Analysis
    - Correlation plots – Time series
    - Density or line charts
  - Multiple Variables
    - Scatter plot
    - Heatmap
    - Stacked bar charts

Improving the data

- Always use business acumen and domain knowledge in creating new variables
  - Convert epoch time in to hh:mm Y:M:D
  - Use epoch time to create a weekday/ weekend column
  - Combine multiple variables to form a new predictive variable
- Adhere to compliance and privacy restrictions of the data set by the region, industry and your company
- Explore other sources of data to supplement your current modeling exercise
  - Use independent weather data from publicly available APIs
  - Use publicly available zipcode information to tie with city and states
  - Use publicly available traffic information to supplement traffic analysis
- If the cardinality(number of unique values) is high for a variable
  - Use the hashing trick
  - Use binning

Simulation work in R

- Quick intro to R
- OOD – NEU Discovery Cluster
- EDA in R
- Break – 5-10 mins

Group Work

- Open Up The Dataset In Course Materials ->Supporting Material->Clean Data
- Go Through The Steps That We Went Through Today
- Present at 8:15
  - What You Cleaned Up
  - Basic Profile Of The Data
  - Any Interesting Insights