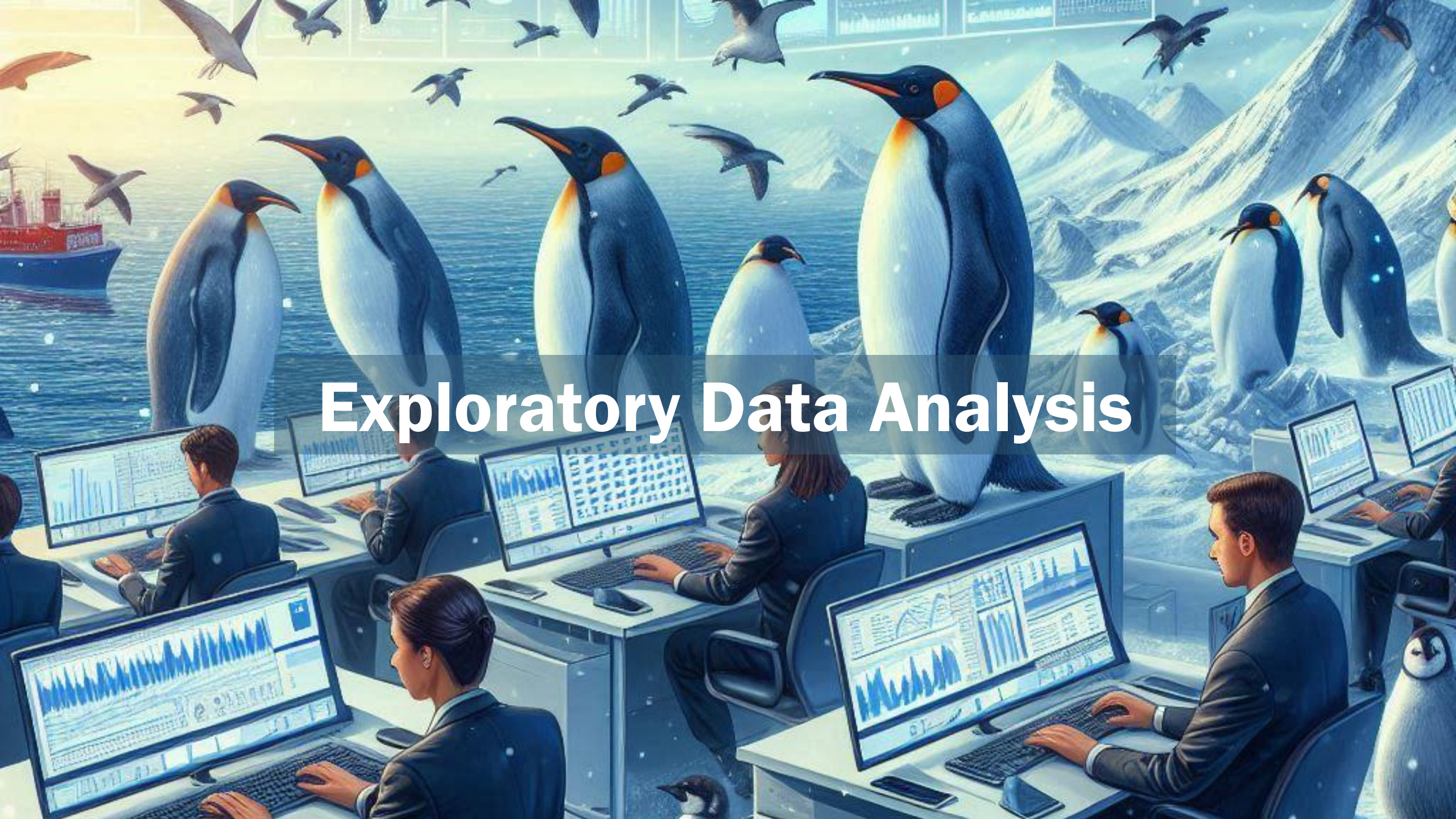
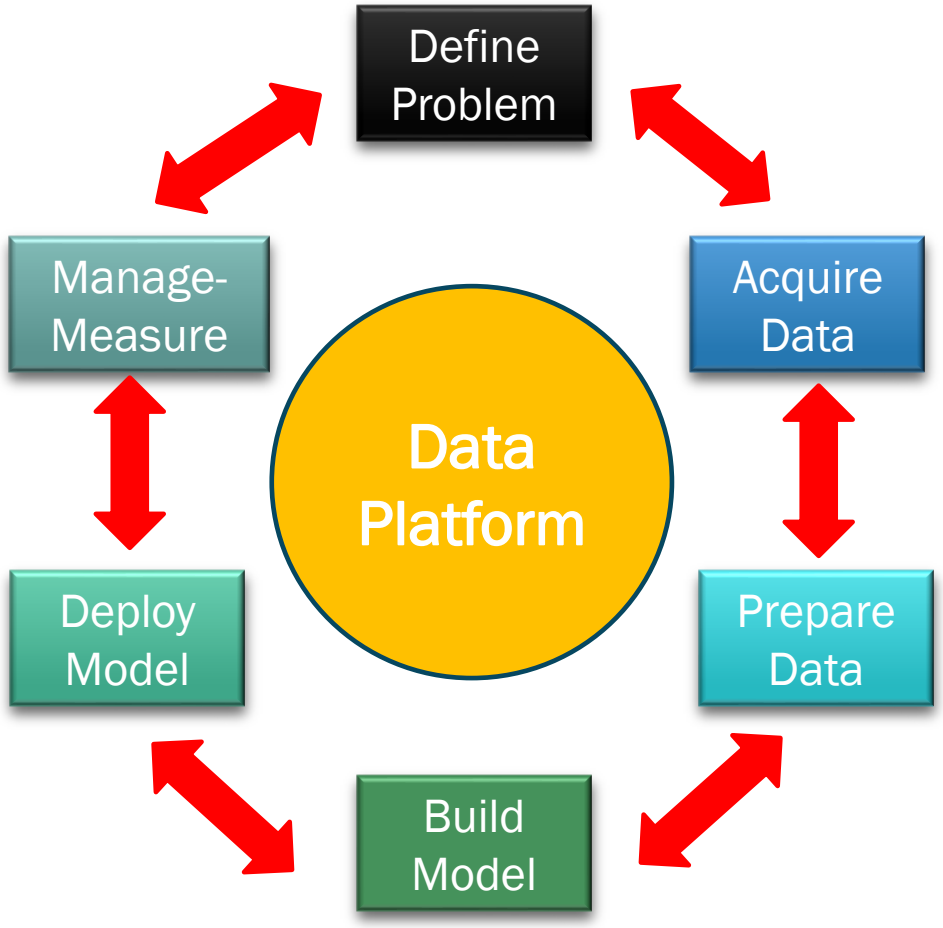

Exploratory Data Analysis

DATA SCIENCE LIFECYCLE



- Identify Data Location
- Manage Metadata
- Extract, Transform, Load
- Join/Merge Tables
- Profile Data
- Data Type and Validation Check
- Regulatory/PII Check

- Clean Data (Missing Values, Outliers, Duplicates, Var names)
- EDA (Univariate, Bivariate, Multivariate analysis, Visualization)
- Prepare Analytic Data Set (ADS) (Standardization/Normalization, Category encoding)
- Feature Engineering

Sometimes the EDA is the objective of the project.

BYU

SUMMARY: DATA ACQUISITION→ PREPARATION→ EXPLORATION

Data Analysis Processes	Tools
Acquire Data	pd.read_csv(), pd.read_excel(), df.to_csv(index = False), df.to_excel(index = False),
Profile Data	df.head(), df.tail(), df.info(), df.describe(include='all'), df.dtypes, df.shape
Manipulate Data	df.groupby(), df.pivot_table(), df.insert()
Clean Data Missing Data Imputing Data Duplicates Outliers Data Types	df.isnull().sum(), df.isna(), df.fillna(value=), df.drop(), df.drop_duplicates(), df.dropna(), df.replace(), df.astype(),
Exploratory Data Analysis Univariate Multivariate Visualization	df.plot(), plt.show(), df.plot.scatter(), df.plot.box(), plt.hist() plt.subplots(figsize=), sns.histplot(), sns.kdeplot(), sns.boxplot(), sns.violinplot() df.mean(), sns.scatterplot(), sns.swarmplot(), sns.stripplot(), plotly.express
Transform Data	preprocessing.scale(), preprocessing.MinMaxScaler().fit_transform() [Part of Modeling because we need to wait until after train/validation/test split]

EXPLORATORY DATA ANALYSIS

Process 4.5.1: Exploratory data analysis.

These four steps should be followed in conducting exploratory data analysis:

Step	Description
Step 1: Understand the data	This is done as part of the ”Profile Data” step before analysis. Objective is to understanding the data types, numbers, ranges, overall cleanliness
Step 2: Detect and address Outliers and missing data	The is done in the Data Cleaning stage. Data needs to be cleaned before analysis; otherwise, analysis could be skewed by dirty data.
Step 3: Describe the shape of each feature of the data	Use descriptive univariate statistics and visualization to characterize data distributions for each feature.
Step 4: Identify and address correlations between features	Use multivariate analysis. Assess whether features with high (+/-) correlations can be dropped.

BYU

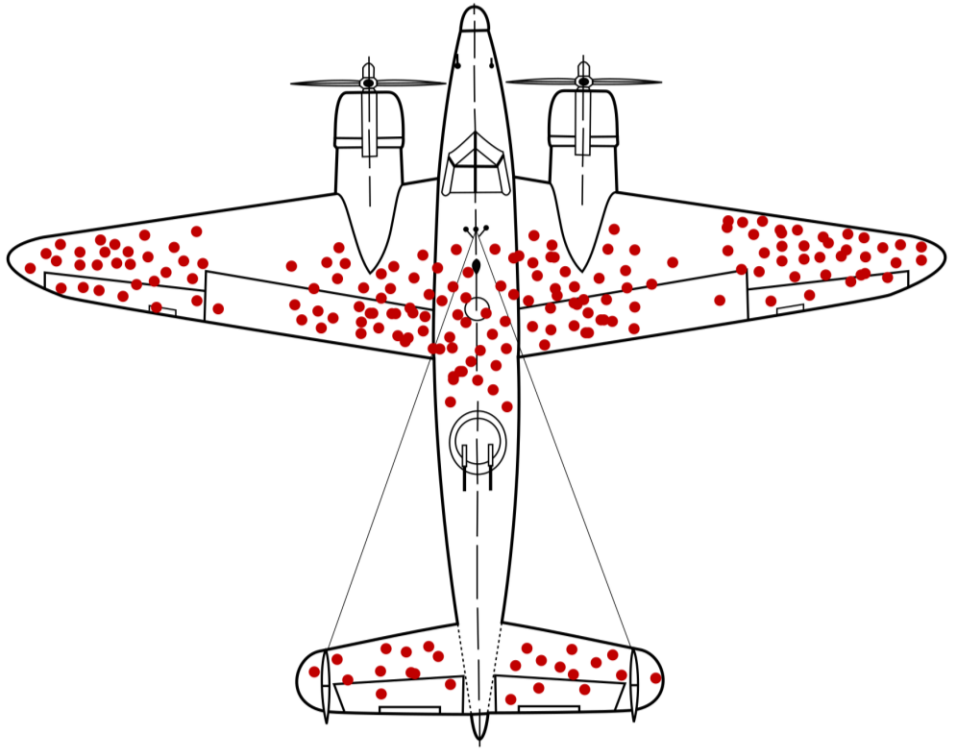
MISSING DATA

"**Missing Completely at Random" (MCAR)** refers to where the missingness occurs purely by chance and has no relationship with any of the variables in the dataset, either the missing ones or the observed ones.

"**Missing at Random" (MAR)** refers to where the missingness can be explained by other known (observed) variables in the dataset, but not by the value of the missing variable.

Missing Not at Random" (MNAR) refers to when the missingness is not random and cannot be explained by the observed data. Instead, it is systematically related to the unobserved, missing values.

Imagine a survey where respondents are asked to provide their age and income, and some survey forms are accidentally lost



me. Some
ne likelihood
ucation level
itself (which

natically less
ariables, the

MISSING DATA QUIZ 1

1. Which of the following best defines “Missing Completely at Random” (MCAR)?

a) The missingness is related to both observed and missing data.
b) The missingness is related to observed data but not missing data.
c) The missingness is unrelated to both observed and missing data.
d) The missingness is related to the missing data only.

2. In which of the following scenarios is data considered “Missing Not at Random” (MNAR)?

a) Data is missing due to a random system failure.
b) Patients with higher incomes are less likely to report their income.
c) Missingness is related to other observed variables, like age or gender.
d) Missing data occurs randomly across the dataset.

3. Which type of missing data is easiest to handle without introducing bias in analyses?

a) Missing Not at Random (MNAR)
b) Missing Completely at Random (MCAR)
c) Missing at Random (MAR)
d) None of the above

4. Which of the following is true about “Missing at Random” (MAR)?

a) Missingness is explained by variables that are not observed.
b) Missingness is random and unrelated to any variables in the dataset.
c) Missingness is related to observed variables but not the missing data itself.
d) Missingness is dependent on the value of the missing data.

BYU

MISSING DATA QUIZ 2

5. Which of the following techniques can help handle data that is “Missing at Random” (MAR)?
a) Removing all missing data rows.
b) Imputing missing values based on observed variables.
c) Ignoring the missing data and proceeding with the analysis.
d) Using the mean of the missing variable to fill in the gaps.

6. If data is Missing Completely at Random (MCAR), removing missing data points will not introduce bias.
(True / False)

7. Data that is Missing Not at Random (MNAR) can be safely ignored without any effect on the analysis.
(True / False)

8. Missing at Random (MAR) means the missingness is related to the unobserved, missing data itself.
(True / False)

BYU

OUTLIER DETECTION

Method	Description
Tukey's Fences	Often used to determine outliers in box plots. 1. Calculate the interquartile range, $\mathbf{IQR} = Q_3 - Q_1$ for a feature. 2. Classify all points that fall 1.5IQR above Q_3 or 1.5IQR below Q_1 as outliers.
z-scores	1. Calculate the z-score $z = \frac{\mathbf{value - mean}}{\mathbf{standard\ deviation}}$ for each value. 2. Classify all points that have a z-score of $\lvert z \rvert > 3$ as outliers.

BYU

MEAN, VARIANCE, STANDARD DEVIATION

The formula for the **population mean** is:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Where:

- μ = population mean,
- X_i = each individual data point in the population,
- N = the total number of data points in the population.

The formula for **population variance** is:

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

Where:

- σ^2 = population variance,
- X_i = each individual data point in the population,
- μ = population mean,
- N = total number of data points in the population.

The formula for **population standard deviation** is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

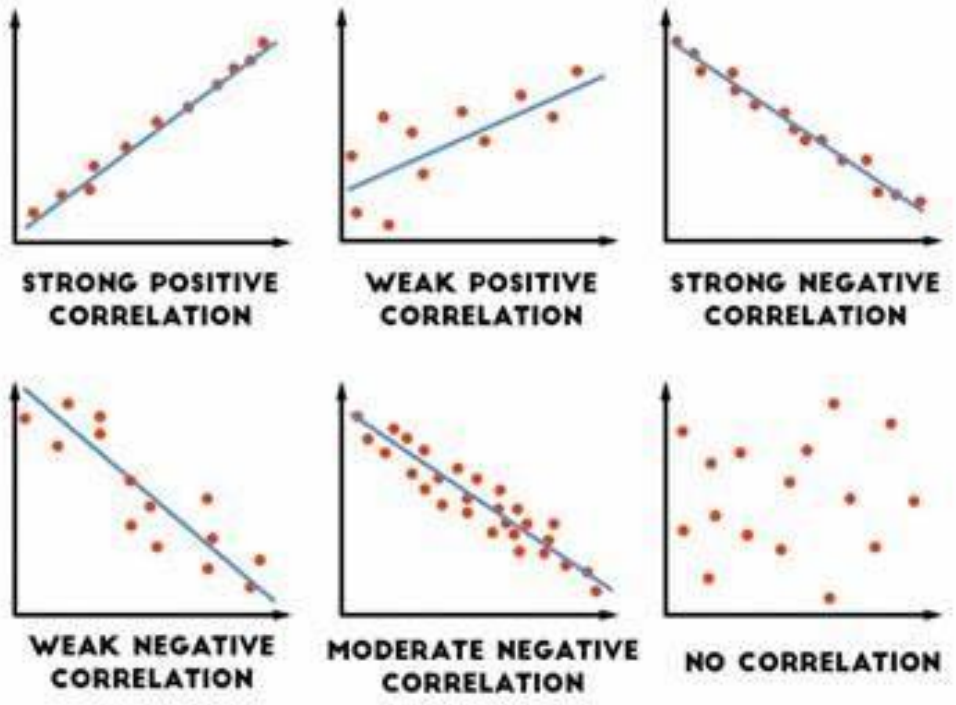
Where:

- σ = population standard deviation,
- σ^2 = population variance,
- X_i = each individual data point in the population,
- μ = population mean,
- N = total number of data points in the population.

BYU

COVARIANCE VS CORRELATION

- **Covariance:** Measures the direction of the relationship between two variables; units depend on the variables.
- **Covariance:** Can range from negative to positive infinity.
- **Correlation:** Measures both direction and strength; ranges from -1 to 1, and is unitless.
- **Correlation:** Normalized, making it easier to compare across datasets.



BYU

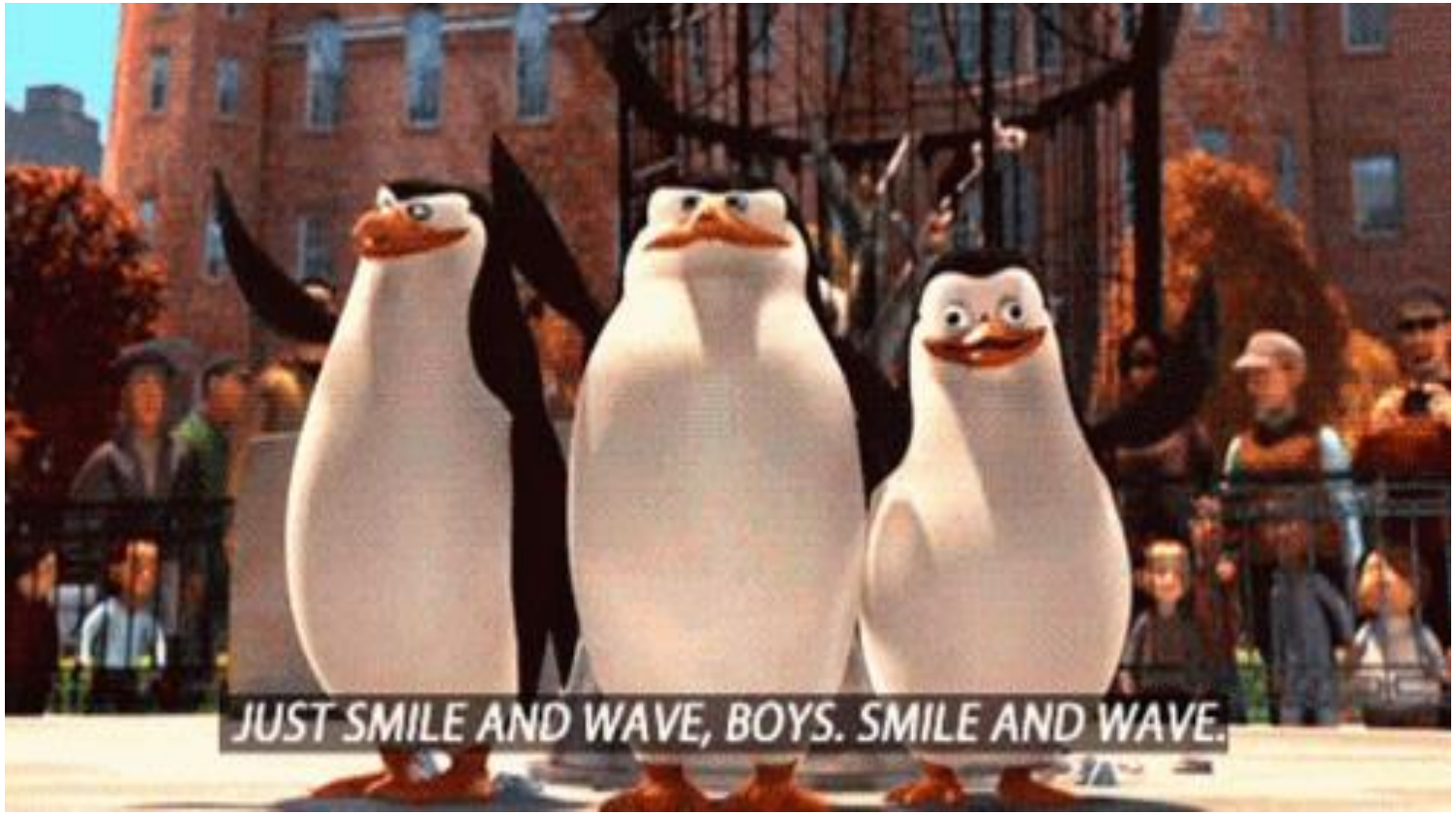
PYTHON CODE

UPDATE LINK:

https://colab.research.google.com/github/rhodes-byu/cs180-winter25/blob/main/notebooks/05a-eda.ipynb

EDA with Palmer Penguins:

https://colab.research.google.com/github/rhodes-byu/cs180-winter25/blob/main/notebooks/05b-eda-penguins.ipynb



BYU