# Stats for Data Science

1. Explain the importance of domain knowledge in data science
   - Being knowledgeable in the field you are working in is crucial. This include understanding the specific challenges and nuances of the field and how to present data effectively.

2. Describe the difference between ratio and interval levels of data measurement. Provide an example for each.
   - **Ratio** level of data – Is a level of measurement where values can be categorized, ordered, have equal interval, and take on a true zero. E.g., you can have no money in the bank, and it make sense that 200000 Is twice as much as 100000.
   - **Interval** level of data – data at interval level allows meaningful subtraction between data points. E.g., temperature – if the temperature Is $32^0$c in Limpopo and $23^o$c in Cape Town, then Limpopo $9^o$c warmer than Cape Town.
3. Explain the process of data cleaning and why it is crucial for data analysis.
   - Data cleaning refers to identifying and correcting errors in data for analysis and visualization. This includes removing duplicates, fill or drop null values, filter data to include necessary fields, and removing the outliers.
   - Data cleaning is crucial for data analysis because we want to analyze data that will give us accurate results to make better data-driven decisions making.
4. A company's workforce productivity is inversely proportional to the number of hours spent in meetings per week. Given the following data:

• When employees spend 4 hours per week in meetings, their productivity is 100 units.
• When employees spend 8 hours per week in meetings, their productivity is 50 units.Using the Information above, answer the following questions:

a) Formulate an equation that links hours spent in meetings (H), productivity (P) and a constant k.

$$P = \frac{k}{H}$$

b) Find the value of k. If the company wants to achieve a productivity of 80 units, how many hours per week should be spent in meetings?

$$P = \frac{k}{H}$$

When employees spend $4_{h/w}$ in meetings, their productivity is 100 units.

$$100 = \frac{k}{4}$$

$$k = 100 * 4$$
$$k = 400$$

$$Therefore\, P = \frac{400}{H}$$

How many hours per week should be spent in meetings to achieve a productivity of 80 units.

$$80 = \frac{400}{H}$$

$$H = \frac{400}{80}$$

$$H = 5$$

To achieve a productivity of 80 units, employees should spend 5 hours per week in meetings.

5. You are analyzing the preferences of a group of customers in a store. The data reveals the following:

· 40% of customers buy Product A.

· 30% of customers buy Product B.

· 20% of customers buy both Product A and Product B.

a) Calculate the probability that a customer buys either Product A or Product B.

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

= 0.4 + 0.3 – 0.2

= 0.5 / 50%

b) Calculate the probability that a customer buys Product A given that they have bought Product B.

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

= 0.2 / 0.3

= 0.6667 / 66.67%

c) Calculate the probability that a customer buys Product B given that they have bought Product A.

$P(B|A) = \frac{P(A \cap B)}{P(A)}$

= 0.2 / 0.4

= 0.5 / 50%

6. Discuss the different types of probability distributions, including both discrete and continuous distributions, and provide examples of each.

- **Normal Distribution**: It is also known as Gaussian distribution. It is one of the simplest types of continuous distribution which is symmetrical around its mean value.
- **Continuous Uniform Distribution** – it is the type of continuous distribution where all outcomes are equally possible; each variable gets the same probability of hit.
- **Pareto Distribution**: It is one of the most critical types of continuous distribution. The Pareto Distribution is a skewed statistical distribution that uses power-law to describe quality control, scientific, social, geophysical, actuarial, and many other types of observable phenomena.
- **Exponential Distribution**: It is a type of continuous distribution that determines the time elapsed between events.
- **Binomial Distribution** - It is one of the popular discrete distributions that determine the probability of x success in the 'n' trial.

- **Geometric Distribution** - is one of the crucial types of discrete distributions that determine the probability of any event having likelihood ‘p’ and will occur after 'n' number of Bernoulli trials.
- **Poisson distribution** - is one of the popular types of discrete distribution that shows how many times an event has the possibility of occurrence in a specific set of time.
- **Multinomial Distribution -** is another popular type of discrete distribution that calculates the outcome of an event having two or more variables.

7. You are provided with a csv file that contains information about a group of people and their respective salaries. The dataset also contains some additional variables. You are required to use it to answer the following questions:
   a) Import the dataset using pandas.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

data = pd.read_csv("salary.csv")
data.head()
```

   b) Using the head() function, view the structure of the data and classify all the columns as either Qualitative or Quantitative then state whether they are nominal, ordinal, interval or ratio.

```
data.head()
Qualitative | Quantitative
 ---------------------------
 - Name      |- Age
 - Gender    |- Salary
 - Position  |- YearAtCompany
 - Department|

 Name -> Nominal level
 Age -> Ratio level
 Gender -> Nominal level
 Position -> Nominal level
 Salary - Ratio level
 Department -> Nominal level
 YearsAtCompany -> Ratio level
```

   c) Provide a brief overview of the ‘Gender’ variable and outline the modal gender and its frequency.

```
gender_counts = data.Gender.value_counts()
print(gender_counts)

# Modal gender and its frequency
modal_gender = gender_counts.idxmax()
```

```
gender_frequency = gender_counts.max()

print(f"Modal Gender: {modal_gender}")
print(f"Frequency: {gender_frequency}")
```

d) Using appropriate visualisations, explore the relationship between 'Education level' and 'Salary'. Explain your findings.

```
plt.figure(figsize=(20, 6))
sns.scatterplot(x='Position', y='Salary', data=data, hue='Position', palette='tab10')
plt.title('Salary Distribution by Position')
plt.ylabel("Salary")
plt.xlabel("Position")
plt.xticks(rotation=90)
plt.show()
```

8. Before writing exams, the IT manager at Richfield Bryanston campus knows that 5% of the computers crash during exams. A quality control inspector randomly selects 20 computers from the lab from the lab to see if it crashes.

a) What is the probability that exactly 2 of the 20 computers selected crash? Show all your workings and round your answer to three decimal places.

b) What is the probability that at most 3 of the 20 cmputers selected crash? Show all your workings and round your answer to three decimal places.

9. Richfield is analyzing the final exam scores of all students who study python. They however discover that the scores are normally distributed with a mean of 68 and a standard deviation of 9.

a) What is the probability that a randomly selected student scored more than 85 on the final exam? Show all your workings and round your answer to four decimal places.

$z = \frac{x-\mu}{\sigma}$

$z = \frac{85-68}{9}$

$z = 17/9$

$z$ = 1.8889 is 0.9706

$P(X > 85) = 1 - 0.9706$

= 0.0294

b) The top 10% of the students will receive an "A" grade. What is the minimum score a student must achieve to be in the top 10% of the class? Show all your workings and round your answer to two decimal places

$z = 1.2816$

$x = \mu + Z * \sigma$

X = 68 + 1.2816 * 9

X = 79.53