**PGPDSE FT Project – EDA**

**STATS**

**Submitted By:** Gayatri Patil

**Dataset and Domain**

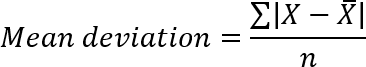
## PART-A (Concept Based)--25 points

The following are the ages of 30 customers who ordered an EV scooter from Zen Automotives.ee.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 42 | 44 | 62 | 35 | 20 |
| 30 | 56 | 20 | 23 | 41 |
| 55 | 22 | 31 | 27 | 66 |
| 21 | 18 | 24 | 42 | 25 |
| 32 | 50 | 31 | 26 | 36 |
| 39 | 40 | 18 | 36 | 22 |

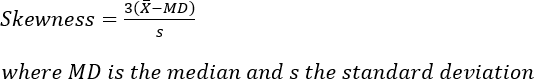
Use this data for answering questions 1-13.

Q1. Compute the mean, median and the mode of the data



Q2. Compute the range, variance and standard deviation of customer ages Q3. Find the mean deviation for the data. The mean deviation is defined as below.

Q4. Calculate the Pearson coefficient of skewness and comment on the skewness of the data

[A measure to determine the skewness of a distribution is called the Pearson coefficient of skewness. The formula is

The value of the coefficient if skewness usually ranges from –3 to 3.

When the distribution is symmetric, the coefficient is zero; when the distribution is positively skewed, the coefficient is positive, and when the distribution is negatively skewed the coefficient is negative.]

Q5. Count the number of data values that fall within two standard deviations of the mean. Compare this with the answer from Chebyshev’s Theorem.

Q6. Find the three quartiles and the interquartile range (IQR).

Q7. Are there any outliers in the data set? Q8. Draw a boxplot of the dataset to confirm.

Q9. Find the percentile rank of the data point 50.

Q10. What is the probability that a person ordering an EV scooter is above 50 years old?

Q11. Create a frequency distribution for the data and visualize it appropriately.

Q12. Create a probability distribution of the data and visualize it appropriately.

Q13. What is the shape of the distribution of this dataset? Create an appropriate graph to determine that. Take 100 random samples with replacement from this dataset of size 5 each. Create a sampling distribution of the mean age of customers. Compare with other sampling distributions of sample size 10, 15, 20, 25, 30. State your observations. Does it corroborate the Central Limit Theorem?

Q14. Treat this dataset as a binomial distribution where *p* is the probability that a person ordering an EV is above 50 years age. What is the probability that out of a random sample of 10 buyers exactly 6 are above 50 years of age?

Q15. A study claims that 10% of all customers for an EV scooter are above 50 years of age. Using the Normal approximation of a Binomial distribution, find the probability that in a random sample of 300 prospective customers exactly 25 will be above 50 years of age.

[Note that the normal distribution can be used to approximate a binomial distribution if np>=5 and nq>=5 with the following correction for continuity P(X=z) = P (z-0.5 < X < z+0.5) ]

Q16. Compute a 95% Confidence Interval for the true mean age of the population of EV scooter buyers for the dataset using appropriate distribution. (State reasons as to why did you use a *z* or *t* distribution)

Q17. A data scientist wants to estimate with 95% confidence the proportion of people who own an EV in the population. A recent study showed that 20% of people interviewed had an EV. The data scientist wants to be accurate within 2% of the true proportion. Find the minimum sample size necessary.

Q18. The same data scientist wants to estimate the proportion of executives who own an EV. She wants to be 90% confident and accurate within 5% of true proportion. Find the minimum sample size necessary.

Q19. A researcher claims that currently 20% of the population are owning EVs. Test his claim with an alpha =0.05 if out of a random sample of 30 two-wheeler owners only 5 own an EV.

Q20. Assume you are working for a Consumer Protection Agency that looks at complaints raised by customers for the transportation industry. Say you have been receiving complaints about the mileage of the latest EV launched by the Zen Auto motives. Zen allows you to test randomly 40 of its new EVs to test mileage. Zen claims that the new EVs get a mileage of 96 kmpl on the highway. Your results show a mean of 91.3 kmpl and a standard deviation of 14.4.

* 1. Show why you support Zen’s claim using the P-value obtained.
  2. After more complaints you decide to test the variability of the mileage on the highway. On questioning Zen’s quality control engineer , you find that they are claiming a standard deviation of 7.2. Test the claim about the standard deviation. [Hint :use the Chi-square test for variance/standard deviation]
  3. Write a summary of results and the action that Zen must take to remedy the complaints.
  4. What is your position on performing the test for variability along with the test for means?

Q21. Write a report comparing and contrasting Descriptive vs Inferential Statistics in about 200-500 words and how each helps to get more insight into the data at hand.

## PART-B (Dataset based)--25 points

* **DOMAIN*:*** Sports
* **CONTEXT*:*** Company X manages the men's top professional basketball division of the American league system. The dataset contains information on all the teams that have participated in all the past tournaments. It has data about how many baskets each team scored, conceded, how many times they came within the first 2 positions, how many tournaments they have qualified, their best position in the past, etc.
* **DATA DESCRIPTION*:*** Basketball.csv - The data set contains information on all the teams so far participated in all the past tournaments.
* ATTRIBUTE INFORMATION:

1. Team: Team’s name
2. Tournament: Number of played tournaments.
3. Score: Team’s score so far.
4. PlayedGames: Games played by the team so far.
5. WonGames: Games won by the team so far.
6. DrawnGames: Games drawn by the team so far.
7. LostGames: Games lost by the team so far.
8. BasketScored: Basket scored by the team so far.
9. BasketGiven: Basket scored against the team so far.
10. TournamentChampion: How many times the team was a champion of the tournaments so far.
11. Runner-up: How many times the team was a runners-up of the tournaments so far.
12. TeamLaunch: Year the team was launched on professional basketball.
13. HighestPositionHeld: Highest position held by the team amongst all the tournaments played.

* **PROJECT OBJECTIVE*:*** Company’s management wants to invest on proposals on managing some of the best teams in the league. The analytics department has been assigned with a task of creating a report on the performance shown by

the teams. Some of the older teams are already in contract with competitors. Hence Company X wants to understand which teams they can approach which will be a deal win for them.

**Steps and tasks: [Total Score: 25 points]**

1. Read the data set, clean the data and prepare a final dataset to be used for analysis.[ 5 points]
2. Perform detailed statistical analysis and EDA using univariate, bi-variate and multivariate EDA techniques to get data driven insights on recommending which teams they can approach which will be a deal win for them. Also as a data

and statistics expert you have to develop a detailed performance report using this data.[ 15 points]

Hint:

Use statistical techniques and visualization techniques to come up with useful metrics and reporting. Find out the best performing team, oldest team, team with highest goals, team with lowest performance etc. and many more.

These are just random examples. Please use your best analytical approach to build this report. You can mix match columns to create new ones which can be used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns. Use graphical interactive libraries to enable you to publish interactive plots in python like plotly.

1. Please include any improvements or suggestions to the association management on quality, quantity, variety, velocity, veracity etc. on the data points collected by the association to perform a better data analysis in future. Submit a 200-500 words report to the management[5-points]

**Business Importance-**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling including machine learning.

**Business Questions Discussion**

**Q1. Compute the mean, median and the mode of the data**

**Code:**

*a=df.mean()*

*a*

*b=df.median()*

*b*

*df.mode()*

**Q2. Compute the range , variance and standard deviation of customer ages.**

**Code:**

df.max()-df.min()

df.var()

c=df.std()

c

**Q3 . Find the mean deviation for the data . The mean deviation is defined as**

**below.**

**Code:**

l1=[]

for i in df:

l1.append(abs(i-a))

print("Mean deviations : " + str(l1))

**Q4. Calculate the Pearson coefficient of skewness and comment on the skewness of the data.**

**Code:**

*sk=3\*(a-b)/c*

*sk*

**Q5. Count the number of data values that fall within two standard deviations of the mean. Compare this with the answer from Chebyshev’s Theorem**

**Code:**

np.ntile()

**Q6. Find the three quartiles and the interquartile range (IQR).**

**Code:**

q1=df.quantile(q=0.25)

q1

q2=df.quantile(q=0.50)

q2

q3=df.quantile(q=0.75)

q3

iqr=q3-q1

iqr

**Q7. Are there any outliers in the data set ?**

**Q8. Draw a boxplot of the dataset to confirm .**

**Code:**

sns.boxplot(df['age'])

**Q8. Find the percentile rank of the datapoint 50.**

**Code:**

df['percentile rank']=df.age.rank(pct=True)

df

df2=df.loc[df['age'] == 50, 'percentile rank'].iloc[0]

df2

**Q9. What is the probability that a person ordering an EV scooter is above 50**

**years old?**

**Code:**

df\_50=len(df[df['age'].astype(int) >= 50])

prob=df\_50/len(df)

prob

**Q10. Create a frequency distribution for the data and visualize it appropriately**

**Code:**

fig, ax = plt.subplots()

df['age'].value\_counts().plot(ax=ax, kind='bar', xlabel='numbers', ylabel='frequency')

plt.show()

**Q11. Create a probability distribution of the data and visualize it appropriately.**

**Code:**

m=np.mean(df)

M

from scipy.stats import poisson

data\_poisson = poisson.rvs(mu=m, size=29)

ax = sns.distplot(data\_poisson,

bins=30,

kde=False,

color='skyblue',

hist\_kws={"linewidth": 15,'alpha':1})

ax.set(xlabel='Poisson Distribution', ylabel='Frequency')

**Q12. What is the shape of the distribution of this dataset? Create an appropriate graph to determine that. Take 100 random samples with replacement from this dataset of size 5 each. Create a sampling distribution of the mean age of customers. Compare with other sampling distributions of sample size 10, 15, 20, 25, 30. State your observations. Does it corroborate the Central Limit Theorem?**

**Code:**

sample\_size = 5 #n=30

n\_smpl=100 #

# create an empty list to store the means of each sample

smpl\_means = []

# create a for loop to get the sample means array

for j in range(n\_smpl):

# generates a random sample of size 100 from the population array

sample = np.random.choice(df['age'], size=j, replace=True)

sample\_mean = np.mean(sample)

# calculate the mean of the sample data

smpl\_means.append(sample\_mean)

# append the mean to the list

sns.distplot(smpl\_means)

plt.title("Sampling Distribution")

plt.show()

**Q13. Treat this dataset as a binomial distribution where p is the probability** **that a person ordering an EV is above 50 years age. What is the probability** **that out of a random sample of 10 buyers exactly 6 are above 50 years of** **age?**

Code:

df\_50=len(df[df['age'].astype(int) >= 50])

prob=df\_50/len(df)

prob

sample\_list10 = random.choices(df['age'], k=10)

print(sample\_list10)

prob1=stats.binom.pmf(k=6,n=10,p=0.1666)

prob1

**Q14. A study claims that 10% of all customers for an EV scooter are above 50 years of age. Using the Normal approximation of a Binomial distribution, find the probability that in a random sample of 300 prospective customers exactly 25 will be above 50 years of age.**

**Code:**

prob2=stats.binom.pmf(k=25,n=300,p=1/10)

prob2

**Q15. Compute a 95% Confidence Interval for the true mean age of the population of EV scooter buyers for the dataset using appropriate distribution.(State reasons as to why did you use a z or t distribution)**

**Code:**

df.mean()

df.std()

mu=34.4666

alpha=0.05

sigma=13.317

n=29

stat, p\_value = stats.shapiro(df['age'])

print('Test statistic:', stat)

print('P-Value:', p\_value)

n=29

t\_critical = round(stats.t.sf( 0.05, df = n-1), 2)

t\_critical

**Q16. A data scientist wants to estimate with 95% confidence the proportion of people who own an EV in the population. A recent study showed that 20% of people interviewed had an EV. The data scientist wants to be accurate within 2% of the true proportion. Find the minimum sample size necessary.**

Code:

n = 100

x = 20

# sample proportion

p\_samp = x/n

interval =stats.norm.interval(0.95, loc=p\_samp, scale=np.sqrt(p\_samp\*(1-p\_samp)/n))

print('95% confidence interval for population proportion is', np.round(interval,2))

**Q17. The same data scientist wants to estimate the proportion of executives** **who own an EV. She wants to be 90% confident and accurate within 5% of**

**true proportion. Find the minimum sample size necessary**

**Code:**

n = 100

x = 20

# sample proportion

p\_samp = x/n

interval =stats.norm.interval(0.90, loc=p\_samp, scale=np.sqrt(p\_samp\*(1-p\_samp)/n))

print('90% confidence interval for population proportion is', np.round(interval,2))

**Q18. A researcher claims that currently 20% of the population are owning**

**#EVs. Test his claim with an alpha =0.05 if out of a random sample of 30**

**#two-wheeler owners only 5 own an EV.**

Code:

n=30

x=5

p\_samp = x/n

interval =stats.norm.interval(0.95, loc=p\_samp, scale=np.sqrt(p\_samp\*(1-p\_samp)/n))

print('90% confidence interval for population proportion is', np.round(interval,2))

**Q19. Assume you are working for a Consumer Protection Agency that looks at complaints raised by customers for the transportation industry. Say you have been receiving complaints about the mileage of the latest EV launched by the Zen Automotives. Zen allows you to test randomly 40 of its new EVs to test mileage. Zen claims that the new EVs get a mileage of 96 kmpl on the**

**highway. Your results show a mean of 91.3 kmpl and a standard deviation of**

**14.4.**

**Code:**

n=40

mu=96

x\_bar=91.3

s=14.4

#a. Show why you support Zen’s claim using the P-value obtained.

z\_stat = (x\_bar - mu) / (s/n\*\*0.5)

z\_stat

p\_value = 2 \* stats.norm.cdf(z\_stat)

p\_value

# p > significace level , therefore accept H0.

# Thus we accept the null hypothesis and conclude that there is not enough evidence to claim that the mean mileage is different from 96kmpl.

#b. After more complaints you decide to test the variability of the mileage on

#the highway. On questioning Zen’s quality control engineer , you find

#that they are claiming a standard deviation of 7.2. Test the claim about

#the standard deviation. [Hint :use the Chi-square test for

#variance/standard deviation]

from scipy.stats import chi2\_contingency

data =df['age']

stat, p, dof, expected = chi2\_contingency(data)

# interpret p-value

alpha = 0.05

print("p value is " + str(p))

if p <= alpha:

print('Dependent (reject H0)')

else:

print('Independent (H0 holds true)')

#c. Write a summary of results and the action that Zen must take to remedy

#the complaints.

**Solution:**

Zen must take a larger sample size and check again the p value if then the null hypothesis

Remains true he must change the claimed mileage.

#d. What is your position on performing the test for variability along with the

#test for means?

**Solution:** performing the test of means on the above dataset concludes that the customer claim stands true i.e. the null hypothesis is accepted and Zen should change the claim of the mileage offered by the EV

**Q20. Write a report comparing and contrasting Descriptive vs. Inferential**

**Statistics in about 200-500 words and how each helps to get more insight into the data at hand.**

**Solution:**

Descriptive Statistics: Descriptive statistics is a term given to the analysis of data that helps to describe, show and summarize data in a meaningful way. It is a simple way to describe our data.

Descriptive statistics is very important to present our raw data ineffective/meaningful way using numerical calculations or graphs or tables. This type of statistics is applied to already known data. Descriptive statistics are used to describe the characteristics or features of a dataset. The term ‘descriptive statistics’ can be used to describe both individual quantitative observations (also known as ‘summary statistics’) as well as the overall process of obtaining insights from these data. We can use descriptive statistics to describe both an entire population and an individual sample. Because they are merely explanatory, descriptive statistics are not heavily concerned with the differences between the two types of data. So what measures do descriptive statistics look at? While there are many, important ones include:

Distribution

Central tendency

Variability

Inferential Statistics: In inferential statistics, predictions are made by taking any group of data in which you are interested. It can be defined as a random sample of data taken from a population to describe and make inferences about the population. Any group of data that includes all the data you are interested in is known as population. It basically allows you to make predictions by taking a small sample instead of working on the whole population. We’ve established that descriptive statistics focus on summarizing the key features of a dataset. Meanwhile, inferential statistics focus on making generalizations about a larger population based on a representative sample of that population. Because inferential statistics focuses on making predictions (rather than stating facts) its results are usually in the form of a probability.

Unsurprisingly, the accuracy of inferential statistics relies heavily on the sample data being both accurate and representative of the larger population. To do this involves obtaining a random sample. If you’ve ever read news coverage of scientific studies, you’ll have come across the term before. The implication is always that random sampling means better results. On the flipside, results that are based on biased or non-random samples are usually thrown out. Random sampling is very important for c.

## PART-B (Dataset Based)

**1. Read the data set, clean the data and prepare a final dataset to be used for analysis.**

**Code:**

df = pd.read\_csv('National\_Stock\_Exchange\_of\_India\_Ltd.csv')

df.head()

interval =stats.norm.interval(0.95, loc=p\_samp, scale=np.sqrt(p\_samp\*(1-p\_samp)/n))

**1.1**

**Code:**

# It can be observed that the data has a lot of values where '-' is used instead of a numeric value.

# since the data is not given, for a better reading in the further analysis we replace the value with np.nan

df = ds.replace('-',np.nan)

df

**1.2**

# In the original data columns with integer values have been given object datatype

# Using the below syntax we change them to their suitable datatype

df[['Score','PlayedGames','WonGames','DrawnGames','LostGames','BasketScored','BasketGiven','TournamentChampion','Runner-up']]=df[['Score','PlayedGames','WonGames','DrawnGames','LostGames','BasketScored','BasketGiven','TournamentChampion','Runner-up']].apply(pd.to\_numeric)

**1.3**

# In the original data columns with integer values have been given object datatype

# Using the below syntax we change them to their suitable datatype

df[['Score','PlayedGames','WonGames','DrawnGames','LostGames','BasketScored','BasketGiven','TournamentChampion','Runner-up']]=df[['Score','PlayedGames','WonGames','DrawnGames','LostGames','BasketScored','BasketGiven','TournamentChampion','Runner-up']].apply(pd.to\_numeric)

**2. Perform detailed statistical analysis and EDA using univariate, bi-variate and multivariate EDA techniques to get data driven insights on recommending which teams they can approach which will be a deal win for them.. Also as a data and statistics expert you have to develop a detailed performance report using this data.**

#### **a. Find the correlation between the columns of the dataset.**

**Code:**

df\_num = df.select\_dtypes(include = np.number)

plt.figure(figsize = (15,8))

sns.heatmap(df\_num.corr(),annot = True)

**b. Teams who scored the highest throughout all the tournaments**

**Code:**

df['Tournament'].unique()sns.heatmap(df\_num.corr(),annot = True)

df.groupby('Team')['Score'].max().head(10).plot(kind = 'bar')

#### **c. Teams who scored the most baskets**

**Code:**

plt.figure(figsize = (15,8))

df.groupby('Team')['BasketScored'].max().head(10).plot(kind = 'pie',autopct = '%1.2f%%')

plt.show()

**d. Teams who gave the most number of baskets in tournaments overall**

**Code:**

df.groupby('Team')['BasketGiven'].max().head(10).plot(kind = 'bar')

**e. relation between played games and wongames by teams**

**Code:**

plt.figure(figsize = (20,20))

df.groupby('Team')['WonGames','PlayedGames'].max().head(10).plot(kind = 'pie',autopct = '%1.2f%%',subplots = True)

plt.show()

### **f. we will now calculate the winning percentage according to the following formula: wins+ 0.5\*ties/total games \* 100%**

**Code:**

### x=0

### df['win%']=0

### for x in range(61):

### df['win%'].iloc[x]=( (df['WonGames'].iloc[x]+ 0.5\* df['DrawnGames'].iloc[x])/ df['PlayedGames'].iloc[x])\*100

### df.replace(np.nan, 0)

#### **g. Teams with lowest performance**

**Code:**

sns.barplot(data = df.nsmallest(10, 'win%'), x='Team', y='LostGames', alpha=0.5)

#### **h.Teams with higest performance**

**Code:**

sns.barplot(data = df.nlargest(10, 'win%'), x='Team', y='LostGames', alpha=0.5)

#### **i. finding outlieallrs**

**Code:**

sns.boxplot(df['WonGames'])

**k. Creating datafarme according to the team won and win percent and plotting the pie chart for teams won the most champion trophies in tournament**

**Code:**

df\_wongames= df.nlargest(20, ['WonGames'])

df\_wongames['TeamLaunch'] = df\_wongames['TeamLaunch'].apply(pd.to\_numeric)

df\_wongames

df\_winpercent = df.nlargest(20, 'win%')

df\_winpercent['TeamLaunch'] = df\_winpercent['TeamLaunch'].apply(pd.to\_numeric)

df\_winpercent

intersected\_df = pd.merge(df\_wongames, df\_winpercent, how='inner')

intersected\_df

plt.pie(intersected\_df['TournamentChampion'].head(), labels= intersected\_df['Team'].head())

plt.title('Teams with most Tournament champion trophies')

plt.show()

**l. top 10 team launch years**

**Code:**

sns.countplot( 'TeamLaunch', data=intersected\_df.head(10))

plt.title('team launch year of the top 10 teams')

plt.show()

intersected\_df[['Team', 'win%', 'WonGames']]

**3. Please include any improvements or suggestions to the association management on quality, quantity, variety, velocity, veracity etc. on the data points collected by the association to perform a better data analysis in future. Submit a 200-500 words report to the management**

# Quality:

#### Some of the columns had mixed data like TeamLaunch, special instructions can be mentioned while adding the data like only enter the year of team formation' to avoid this.

# Quantity:

#### Since our dataset had only 61 entries, we couldn't figure out more trends. We can include information about the players of each team, at least of the captains. It would have given us some information about the team's performance under different captains.

# Velocity:

# Efficient speed of data collection can be achieved if we maintain a regular record of the games, keep adding data after every match.

# Variety:

# Again, more information about the team players can help us add variety to the data and will help us understand trends over the years.

# Conclusion:

#### Company X can approach the top 5 teams namely Team 1,2,3,4,5 since they have the highest score, max no of baskets scored, max winning % and most.