Exploratory Data Analysis and visualization

Mohana Dharshini
23BAD066

UNIT - 1

Analysis Techniques - Assignment

Question - 1

Scenario : Online Food Delivery performance.

A food delivery company recorded delivary times (in min) for one hour.

order ID	Deliver time (min)
01	22
02	24
03	21
04	23
05	95

tasks:

a) Identify the level of measurement of the variable.

Ratio Scale

* Differences and ratios are meaningful.
* Because delivery time has a true zero (0 min mea

no time)

b) Compute mean and median

* Mean = $\frac{21+28+23+24+95}{5} = 37$

* Median (Middle value) = 23 (23 min)

c) Identify if an outlier exists

* 95 min is an outlier
* It is far away from the other delivery times (21-28)

d) State which measure best represents typical delivery times and justify.

* Median (23 min) best represents typical delivery time

* Mean (37) is pulled upward due to the outlier (9
* Median is not affected by extreme values.

Question : 2 : Student feedback Analysis

Scenario : Student feedback Analysis:

A college collected course feedback ratings.

Student	Rating (1-5)
S_1	5
S_2	4
S_3	5
S_4	2
S_5	1

Tasks:

a) Identify the level of Measurement.

Ordinal Scale

* Ratings (1-5) show order
* But the gap between 1 and 2 may not equal the gap between 4 and 5.

---

b) Compute mode and median:

* Mode (Most frequent value) : 5
* Median = 4

---

c) Is mean appropriate? explain.

* Mean is not the best choice for ordinal data.
* Ratings are categories with order, not exact numeric distances.
* Mean assumes equal intervals between values, which may not be true.

d) Suggest a Suitable visualization.

Bar Chart (Frequency Ratings)

X-axis : Rating (1 to 5)

Y-axis : Number of Students

Question : 3

Scenario : Temperature and Equipment Failure.

A factor records machine failures at different temperatu

Day	Temperature(°C)	Failures
D_1	35	0
D_2	32	1
D_3	30	2
D_4	28	4
D_5	25	6.

tasks :

a) Identify independent and dependent variables.

*Independent variable : temperature (°C)

*Dependent variable : Number of failures

---

b) Compute mean temperature and mean failures.

*Mean temperature : $\frac{35+32+30+28+25}{5} = \frac{150}{5} = 30$

*Mean Failures = $\frac{0+1+2+4+6}{5} = \frac{13}{5} = 2.6$

c) Identify the trend

* Negative relationship (inverse trend).
* As temperature decreases, failures increase.

---

d) Comment on risk as temperature decreases.

* Risk increases as temperature drops.
* At 35°C → 0 failures
* At 25°C → 6 failures. So machines are more likely to fail in colder temperatures, meaning higher breakdown risk.

Question : 4

Scenario : Data cleaning in Banking

A bank records transaction amounts (₹)

Transaction ID	Amount
T_1	5000
T_2	7000
T_3	-300
T_4	7000
T_5	(blank)

tasks :

a) Identify data quality issues :

1. Negative amount (-300) → may be invalid.

2. Missing value (blank) for T_5

3. Duplicate amount (7000 repeated) → may be valid but needs checking.

4. Possible data entry error / inconsistant recording

---

b) Suggest cleaning actions

1) Check -300

* If refund allowed → keep it and label as refund.

* if not allowed → correct / remove after verification

2) Handle blank value.

* Fill using correct source record (bank log)

* if not available → mark as missing (Null) or remove row.

3) Verify duplicate amounts.

* Ensure T_2 and T_5 are real transactions, not duplicate entry.

4) Add validation rules:

* Amount should be numeric

* No blanks allowed

* Negative only allowed for refund type.

c) State the impact of not cleaning this data.

If not cleaned, it can cause:

* wrong total transaction Value.
* Wrong Customer balance reports.
* Incorrect fraud detection results
* Bad business decisions (profit/loss)
* Issues in auditing and compliance.

Question : 5 : Website traffic

Scenario : Website traffic

Daily visitors recorded for a website

Day	visitors
Mon	120
tues	130
Wed	125
Thu	500
Fri	135

task:

a) Identify the outlier.

* Thursday = 500 visitors is the outlier. (because others are around 120 - 135)

---

b) Compute Mean and Median

Mean = $\frac{120+130+125+500+135}{5} = \frac{1010}{5} = 202$

Median = 130

c) Which value is more representative and why?

* Median (130) is more representative.
* Mean (202) is inflated because of the outlier (150
* Median shows the normal daily traffic level.

d) What could cause such an outlier?

Possible reasons for 1500 visitors?

* Special offer / discount compaign, viral social me post, influencer promotion, Festival / holiday traffic sp email marketing compaign, website featured in news trending, Bot traffic / Spam visits.

Question 6 : Hospital patient Monitoring :

A hospital monitors patient heart rate.

patient	Heart Rate (bpm)
P1	72
P2	75
P3	70
P4	180
P5	74
P6	73.

tasks :

a) Identify the level of measurement

Ratio Scale

* Heart rate has a true zero (0 bpm = no heart beat)
* Differences and ratios are meaningful.

b) Compute mean, median and range.

Mean: $\frac{72+75+70+180+74+73}{6} = \frac{544}{6} = 90.67$

Median: average of 3rd and 4th values.

$\frac{73+74}{2} = 73.5$ Range = 180 - 70 = 110

c) Identify outliers:

180 bpm (patient P_4) is an outlier because it is far higher than normal resting heart rates (around 60 - 100 bpm)

d) Interpret ~~clinical~~ clinical risk.

* P_4 = 180 bpm indicates high clinical risk.
* Severe tachycardia, Stress / panic attack, Fever / infection, Heart rhythm problem (arrhythmia), Emergency Condition needing immediate attention.

e) Why EDA is critical before prediction.

* Detect outliers, identify data errors or abnormal reading Understand distribution, choose correct model, Improve prediction accuracy by cleaning and preprocessing

Question: 7 : Sales data for Retail chain.

Monthly Sales (₹) of a Store.

Months	Jan	Feb	Mar	Apr	May	Jun
Sales	25	27	26	28	90	29

tasks:

a) Identify measurement Scale.

Ratio Scale: Sales have true Zero (0 sales = no revenue)
Ratios make Sense (90 is 3 x of 30)

b) Compute mean, median, Standard deviation.

Mean = 37.5, Median = $\frac{27+28}{2}$ = 27.5,

SD & variance = $\frac{3317.5}{6}$ = 552.92 SD = $\sqrt{552.92}$

SD = 23.52 lakhs

c) Detect outliers

May = 90 lakhs is an outlier

d) How outlier affects decision-making

* Mean becomes very high (37.5) even through normal Sales are 27-29.
* company may assume Store is performing better than reality.
* Wrong Inventory planning.
* Wrong Sales targets.
* Incorrect budgeting and forecasting.