Maryam Soleimani 401222075

1 Theoretical Questions

Question 1:

Explain the difference between correlation and causation with an example from a real-world dataset.

Causation means action A causes outcome B.

On the other hand, correlation is simply a relationship where action A relates to action B—but one event doesn't necessarily cause the other event to happen.

Example:

- **Dataset Observation:** In summer months, both ice cream sales and drowning deaths increase.
- **Correlation:** There is a positive correlation between ice cream sales and drowning incidents.
- **Misinterpretation :**"Eating more ice cream causes drowning."
- **Actual Causation:** The real cause is **hot weather ,** people swim more (increasing drowning risk) and eat more ice cream.

Question 2:

(a) What are the major types of issues found in raw data, and how do they affect analysis?

(b) List the four major tasks of data preprocessing.

(c) What are some methods for handling missing values in a dataset?

a)

1. Inaccurate Data

Inaccurate data is data that is wrong—customer addresses with the wrong ZIP codes, misspelled customer names, or entries marred by simple human errors. Whatever the cause, whatever the issue, inaccurate data is unusable data. If you try to use it, it can throw off your entire analysis.

2. Incomplete Data

Another common data quality is incomplete data. These are data records with missing information in key fields—addresses with no ZIP codes, phone numbers without area codes, and demographic information without age or gender entered.

Incomplete data can result in flawed analysis. It can also make daily operations more problematic, as staff scurries to determine what data is missing and what it was supposed to be.

3. Duplicate Data

When importing data from multiple sources, it's not uncommon to end up with duplicate data. For example, if you're importing customer lists from two sources, you may find several people who were customers of both retailers. You only want to count each customer once, which makes duplicative records a major issue.

4. Inconsistent Formatting

Much data can be formatted in a multitude of ways. Consider, for example, the many ways you can express a date—June 5, 2023, 6/5/2023, 6-5-23, or, in a less-structured format, the fifth of June, 2023. Different sources often use different formatting, so these inconsistencies can result in major data quality issues.

Working with different forms of measurement can cause similar issues. If one source uses metric measurements and another feet and inches, you must settle on an internal standard and ensure that all imported data correctly converts.

5. Cross-System Inconsistencies

Inconsistent formatting is often the result of combining data from two different systems. It's common for two otherwise-similar systems to format data differently. These cross-system inconsistencies can cause major data quality issues if not identified and rectified.

6. Unstructured Data

While much of the third-party data you ingest will not conform to your standardized formatting, that might not be the worst problem you encounter. Some of the data you ingest may not be formatted at all.

This unstructured data can contain valuable insights but doesn't easily fit into most established systems

7. Dark Data

Hidden data, sometimes known as dark data, is collected and stored by an organization that is not actively used

In many cases, it's a wasted resource that many organizations don't even know exists—even though it can account for more than half of an average organization's data storage costs.

8. Orphaned Data

Orphaned data isn't hidden. It's simply not readily usable. In most instances, data is orphaned when it's not fully compatible with an existing system or not easily converted into a usable format. For example, a customer record that exists in one database but not in another could be classified as an orphan.

Data quality management software should be able to identify orphaned data. Thus identified, the cause of the inconsistency can be determined and, in many instances, rectified for full utilization of the orphaned data.

9. Stale Data

Data does not always age well. Old data becomes stale data that is more likely to be inaccurate. Consider customer addresses, for example. People today are increasingly mobile, meaning that addresses collected more than a few years previous are likely to reflect where customers used to live, not where they currently reside.

Older data needs constant culling from your system to mitigate this issue. It's often easier and cheaper to delete data past a certain expiration date than to deal with the data quality issues of using that stale data.

10. Irrelevant Data

Many companies capture reams of data about each customer and every transaction. Not all of this data is immediately useful. Some of this data is ultimately irrelevant to the company's business.

Capturing and storing irrelevant data increases an organization's security and privacy risks. It's better to keep only that data of immediate use to your company and either delete or not collect in the first place data of which you have little or no use.

b)

data cleaning, data integration, data reduction, and data transformation

c)

- **Deletion:** This involves removing rows or columns with missing values. This is a straightforward method, but it can be problematic if a significant portion of your data is missing. Discarding too much data can affect the reliability of your conclusions.
- **Imputation:** This replaces missing values with estimates. There are various imputation techniques, each with its strengths and weaknesses. Here are some common ones:
    - **Mean/Median/Mode Imputation:** Replace missing entries with the average (mean), middle value (median), or most frequent value (mode) of the corresponding column. This is a quick and easy approach, but it can introduce bias if the missing data is not randomly distributed.
    - **K-Nearest Neighbors (KNN Imputation):** This method finds the closest data points (neighbors) based on available features and uses their values to estimate the missing value. KNN is useful when you have a lot of data and the missing values are scattered.
    - **Model-based Imputation:** This involves creating a statistical model to predict the missing values based on other features in the data. This can be a powerful technique, but it requires more expertise and can be computationally

Question 3:

Describe the 'binning' method for managing noisy data. Give an example.

Noise is a random error or variance in a measured variable.

Binning method: Binning methods smooth a sorted data value by consulting its "neighborhood," i.e. the values around it. The sorted values are distributed into a number of "buckets," or bins.

Binning Methods for Data Smoothing

Let's look at the following data smoothing techniques:

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

*** Partition into (equal frequency) bins:**

- Bin 1: 4, 8, 9, 15

- Bin 2: 21, 21, 24, 25

- Bin 3: 26, 28, 29, 34

*** Smoothing by bin means:**

- Bin 1: 9, 9, 9, 9

- Bin 2: 23, 23, 23, 23

- Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:

- Bin 1: 4, 4, 4, 15

- Bin 2: 21, 21, 25, 25

- Bin 3: 26, 26, 26, 34

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9.

Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median.

In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing.

Question 4:

(a) Discuss the importance of data quality in EDA, including common issues like outliers and inconsistencies.

(b) Describe a scenario where data quality issues might lead to misleading conclusions in a real-world analysis.

(c) Explain how EDA techniques can be used to identify and address these issues.

a)

Exploratory Data Analysis (EDA) is a crucial step in data science that helps uncover patterns, detect anomalies, and guide modeling decisions. However, the effectiveness of EDA heavily depends on data quality. Poor-quality data can lead to misleading insights, flawed models, and incorrect business decisions.

Why Data Quality Matters in EDA

1. Reliable Insights – High-quality data ensures that trends, correlations, and statistical summaries accurately represent the underlying reality.
2. Better Decision-Making – Businesses rely on EDA for strategic decisions; bad data leads to poor choices.
3. Model Performance – Machine learning models trained on flawed data will perform poorly in real-world scenarios.
4. Efficiency – Cleaning data early (during EDA) saves time and effort in later stages.

Outliers: Extreme values that deviate significantly from the rest of the data. they Distorts statistical measures (mean, variance) and misleads models. We can detect them by Visualization: Box plots, scatter plots, histograms and Statistical methods: Z-scores, IQR (Interquartile Range).and we can solve them by removing them or Transform (log, scaling) or cap/winsorize extreme values.

Inconsistencies: Data entries that don't follow expected formats .it Causes errors in aggregation and analysis.

We can detect them by Frequency counts of categorical variables or Regular expressions for text validation.

We can solve them by Standardize formats (convert all country names to a single standard) or Use data validation rules (enforcing date formats).

b)

for example considering that a shop wants to find out why customers stop ordering. They look at their data and see: Some customers have no order history (blank in the data). The Wrong Conclusion can be: "New customers leave quickly!" but the Truth is: The missing data was from a system error—many were loyal customers. This issue was caused by data missing. Another issue can be Wrong Numbers: A typo says one customer ordered 100 pizzas (real max: 5).Wrong Conclusion: "Big orders mean happier customers!"Truth: The extreme number messed up the average. And another example of data quality issue can be: Mixed-Up Labels. Some orders are marked "delivered," others "completed." Wrong Conclusion: "Delivery doesn't affect satisfaction!" Truth: They're the same thing—bad labeling hid the real trend.

C)

There are four steps: data profiling, data cleaning, data transformation, and data validation.

1Data profiling

Data profiling is the first step of EDA, where you collect basic information about your data, such as the number of rows and columns, the data types, the range of values, the missing values, the outliers, the duplicates, and the

inconsistencies. Data profiling helps you understand the structure, content, and quality of your data, and identify potential problems that need further investigation or correction. You can use various tools and techniques for data profiling, such as descriptive statistics, frequency tables, histograms, box plots, scatter plots, and correlation matrices.

2Data cleaning

Data cleaning is the process of fixing or removing data quality issues that you identified in the data profiling step. Data cleaning can involve various tasks, such as removing or imputing missing values, removing or adjusting outliers, removing or merging duplicates, correcting or standardizing values, and resolving or documenting inconsistencies. Data cleaning can improve the accuracy, completeness, consistency, and integrity of your data, and make it ready for further analysis. You can use various tools and techniques for data cleaning, such as conditional statements, functions, formulas, filters, sorting, grouping, and aggregation.

3Data transformation

Data transformation is the process of modifying or creating new variables from your existing data, to make it more suitable for your analysis goals. Data transformation can involve various tasks, such as scaling, normalizing, or standardizing values, creating or deleting variables, splitting or combining variables, encoding or decoding variables, and applying mathematical or logical operations. Data transformation can enhance the meaning, relevance, and usability of your data, and make it easier to apply analytical methods and models. You can use various tools and techniques for data transformation, such as arithmetic operations, logical operations, string operations, date and time operations, and functions.

4Data validation

Data validation is the final step of EDA, where you check and confirm that your data is accurate, complete, consistent, and relevant for your analysis goals. Data validation can involve various tasks, such as verifying the data sources, checking the data assumptions, testing the data hypotheses, comparing the data results, and evaluating the data quality. Data validation helps you ensure that your data is trustworthy, reliable, and valid, and that your analysis results are meaningful, relevant, and actionable. You can use various tools and techniques for data validation, such as cross-validation, error analysis, sensitivity analysis, performance metrics, and visualization.

Question 5:

What is normalization, and why is it important in EDA? Name three methods.

It is an important aspect of data management and analysis that plays a crucial role in both data storage and data analysis. It is a systematic **approach to decompose data tables to eliminate redundant data** and undesirable characteristics.

The primary goal of data normalization is to add, delete, and modify data without causing data inconsistencies. It ensures that **each data item is stored in only one place** which reduces the overall disk space requirement and improves the consistency and reliability of the system.

Data normalization has applications in a wide array of fields and professions. Its ability to streamline data storage, reduce data input error, and ensure consistency makes it an invaluable asset for anyone dealing with large datasets. Let’s discuss some of its use cases.

Decimal place normalization

Decimal place normalization occurs in data tables with numerical data types. If you've ever played with Excel, you know how this happens. By default, Excel places two digits after the decimal for normal comma-separated numbers. You have to decide how many decimals you want, and scale this throughout the table.

Data type normalization

Another common type of normalization is data *types*, and more specifically, *subtypes of numerical data*. When you build a data table in Excel or a SQL-queried database, you may find yourself looking at numerical data that's sometimes recognized as a currency, sometimes as an accounting number, sometimes as text, sometimes as general, sometimes as a number, and sometimes as comma-style. As a list, the data type possibilities for numbers are:

- Currency
- Accounting number
- Text
- General
- Number
- Comma-style

The problem is that these subtypes of numerical data respond differently to formulas and various analytical treatments. In other words, you need them to be normalized to the same type.

In my experience, **the best type to reference by default is comma-style**. It's the easiest to read, and can be labeled as a currency or accounting number if a presentation is later. Moreover, it undergoes the fewest updates over time, so your Excel file stays relevant across programs and across time.

Formatting normalization

A final easy normalization technique is formatting. While this is most common for strings (text, not numbers), it can happen with numbers as well. In most cases of formatting inconsistencies, the challenge is with ***italics***.

This is an easy fix, and most digital natives are familiar with it. Just highlight your italicized, bolded, or underlined cells and navigate to **Home>Bold/Italics/Underline**.

Having one of these typographical emphases will not disturb your analysis. However, it can be distracting and prevent you from catching more significant inconsistencies, such as decimal places and data types.

Question 6:

What is the goal of data reduction, and what techniques are commonly used

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but still contain critical information. Data reduction can be achieved several ways. The main types are data deduplication, compression and single-instance storage.

Techniques:

- Dimensionality Reduction: This approach attempts to reduce the number of “dimensions,” or aspects/variables, from a data set. For example, a spreadsheet with 10,000 rows but only one column is much simpler to process than one with an additional 500 columns of attributes included. This approach can include compression transformations or even the removal of irrelevant attributes for a specific data mining application.
- Data Cube Aggregation: This technique aggregates multidimensional data at various levels to create a “data cube” (or multidimensional data object). This simply means that the data is processed and reduced to a smaller but equally useful form of information that still represents trends in relevant information for analytics.
- Numerosity Reduction: Numerosity reduction, as the name suggests, is replacing the original data with a smaller form of data representing the original with more or less fidelity. This common compression technique is used in various data types, including audio, video, and image.
- Clustering: This practice uses data attributes to build a set of clusters in which the data is split. Similarities and dissimilarities between the data objects result in different placements and distances between the clusters and the objects as a whole.

Question 7:

(a) Why is data visualization considered a powerful storytelling tool in data science?

(b) Compare traditional data presentation with storytelling visualizations using an example. What design elements transform a chart into a compelling story?

a)

Data visualization acts as a powerful tool by using charts, graphs and other visuals to not only present information but also guide viewers through a journey of discovery. This journey encourages curiosity, fosters understanding and encourages action.

b) Data visualization is simply the visual representation of data. This might be basic charts and tables that are generated from a spreadsheet. Or, it could go well beyond those modalities to include any use of shapes, color, and sizing to draw visual focus to data findings. Bottom line, data visualization is about communicating the substance of your metrics in a visual way.

Data visualization can certainly be used to tell a story at the slide level. It can:

- Provide context
- Elevate and draw attention to key insights (and visually subdue the others)
- Lead to action (AKA: the “ask”)

Storytelling with data differs from data visualization because it requires communicators to offer a larger, holistic, view of their message. You must focus first on your audience and structure a larger message before any visuals are rendered. You must identify from the start:

- What do I want my audience to *know* or *do* with the data I am presenting?
- How will I structure a narrative that leads to desired action?
- How is my data helping drive a decision?

There is no understating how important it is for all presented data to have a purpose. Every piece of data you include should further this purpose – or it should be left out.

1. Understand your audience

Audience understanding – in the form of targeting and segmentation – is vital in data storytelling, allowing marketers to tailor a narrative to a specific group and increase its impact in the process.

It's common sense: by understanding the characteristics, needs, and preferences of different audience segments, data storytellers can craft messages that resonate and drive engagement.

The result supports a stronger connection between brand and audience because it feels personal and relevant. If a data story doesn't feel relevant, then by definition it's irrelevant – and that's not exactly the ideal place to start.

2. Gather your data

If understanding your audience is the first step to great data storytelling, the next is actually gathering the appropriate data.

Think of individual data points as pigments on an artist's palette – collecting more and better data enables storytellers to paint a far richer picture of consumer behavior and market trends as they craft stories that resonate and influence audience actions.

3. Turn your data into insights

Data analysis is a key ingredient in the data storytelling recipe. It's the process of sifting through masses of information to identify patterns and trends that could otherwise go unnoticed. The result of this process are *insights* – raw data that's presented in a way that gives it context and meaning.

Without this analytical deep dive, a data story could be just a collection of facts that don't add up to anything in particular. Analysis transforms isolated facts into a coherent narrative that highlights what's important, revealing the story behind the numbers and enabling audiences to make informed decisions.

4. Visualize your data story

Imaginative data visualization and representation make a data story digestible, highlighting patterns, trends, and insights that might not be obvious in, say, a table of figures.

By transforming numbers into visuals, data stories immediately become accessible, engaging, and memorable.

Visualization brings data to life and gives it a voice so it speaks directly to its audience. Whether using graphs, charts, or infographics, visualization is crucial for crafting stories that resonate in a way that raw data simply can't.

5. Build your whole story around data

If you're going to use data storytelling, *really use it*. Embrace it, commit to it. Make data integral to your story, and you'll be rewarded. Tacking a few insights onto an existing narrative and hoping the result will work is unlikely to succeed.

As we've said, data storytelling is about turning abstract numbers into relatable stories, making complex information digestible and engaging for audiences in a way that lends credibility and authority to the message. It's a way to guide audiences through a journey of understanding, where each data point is a landmark, making the data story not just interesting but also persuasive. And the full benefit of that only comes when data is integral.

6. Keep your story simple, relatable, and interesting

The core message of your data story should be accessible, relevant, and easy to explain; that's the best way to form a deep and genuine connection with consumers.

Data storytellers have a tiny window of opportunity to engage audiences with content, so make sure what you have to say is concise, impactful, and engaging.

How exactly? Use audience data to understand what content they're expecting to see on what channel, and then make sure you give it to them. Get that right and you'll…

7. Tell stories that readers will want to share

Word-of-mouth is one of the most powerful marketing tools, and in today's social media-saturated world, a good data story can be shared millions of times each day on multiple platforms.

To make sure your story is one of them, it's important to gather insights from other brands your audience follows to get a better understanding of what stories are being retold, how these can inform your brand's message, and how you can take a fresh approach. Make it share-worthy and there's every chance people will share it. The opposite is also true.

Question 8:

(a) What factors determine the best type of chart to use for a dataset?

(b) How can distribution charts help in EDA, and what questions do they answer?

(c) Discuss how a heatmap of a correlation matrix can help identify patterns in a multivariate dataset.

a) **1. What's Your Goal?**

- **Compare values?** → *Bar chart* (e.g., sales by product).
- **Show trends over time?** → *Line chart* (e.g., monthly revenue).
- **Reveal distributions?** → *Histogram/Box plot* (e.g., age groups).
- **Display proportions?** → *Pie chart* (if few categories) or *Stacked bar*.
- **Show relationships?** → *Scatter plot* (e.g., height vs. weight).

2. What's Your Data Type?

- **Categorical data** (e.g., cities, product names) → *Bar charts*.
- **Numerical data** (e.g., sales, temperature) → *Line/Scatter plots*.
- **Geospatial data** (e.g., country stats) → *Maps/Heatmaps*.

3. Who's Your Audience?

- **Executives?** → Simple, high-impact charts (e.g., *trend lines*).
- **Analysts?** → Detailed charts (e.g., *box plots*).
- **General public?** → Visuals with minimal jargon (e.g., *pictograms*).

b)

1. **"What's the typical value?"**
    - Shows where most data points cluster (e.g., peak of a histogram).
    - *Example:* A histogram of customer ages might reveal most users are 25–34.
2. **"How spread out is the data?"**
    - Identifies variability (tight vs. wide distributions).
    - *Example:* A narrow box plot of exam scores means most students scored similarly.
3. **"Are there outliers or weird patterns?"**
    - Flags anomalies (e.g., extreme values, gaps, or bimodal distributions).
    - *Example:* A long tail in a histogram could indicate rare but high-spending customers.
4. **"Is the data symmetric or skewed?"**
    - Left/right skew impacts statistical models.
    - *Example:* Income data is often right-skewed (few very high earners).

5 "Does it match expected patterns?"

- Checks for normality (bell curve) or other expected shapes.
- *Example:* If website load times aren't log-normal, there might be technical issues.

c)

Heatmaps excel at visualizing the correlation matrix between multiple variables, making it easy to identify highly correlated or inversely correlated variables at a glance. Heatmaps are also useful for visually comparing data across two dimensions, such as different time periods or categories.

1. Visualizing Pairwise Relationships

- A correlation matrix quantifies the linear relationship between all pairs of numerical variables (typically using **Pearson's r**).
- The heatmap color-codes these correlations, making it easy to spot strong positive (near +1), strong negative (near -1), and weak (near 0) relationships at a glance.

2. Identifying Clusters of Correlated Variables

- Variables that are highly correlated with each other often appear as **blocks of similar colors** in the heatmap.
- This helps in detecting **redundant variables** (e.g., multicollinearity in regression models) or grouping related features (e.g., in feature selection).

3. Revealing Negative Correlations

- Strong negative correlations (dark colors in an opposite scale) indicate inverse relationships, which can be meaningful in certain analyses (e.g., one variable increases while the other decreases).

4. Guiding Feature Selection & Dimensionality Reduction

- If two variables are **highly correlated**, one may be dropped to reduce redundancy (e.g., before applying PCA or machine learning).
- Helps in selecting a subset of variables that capture the most variance.

5. Detecting Hidden Patterns & Anomalies

- Unexpected correlations may reveal **hidden relationships** worth further investigation.
- Weak correlations may indicate **independent variables**, useful for assumptions in statistical modeling (e.g., linear regression assumes low multicollinearity).

6. Enhancing Interpretability with Dendrograms (Optional)

- When combined with **hierarchical clustering**, the heatmap can group similar variables together, making patterns even clearer.

A correlation heatmap provides an **intuitive, compact summary** of relationships in multivariate data, aiding in exploratory analysis, modeling, and hypothesis generation.

Question 9:

Compare bar charts, line charts, and pie charts in terms of the types of insights they provide and the nature of the data they are best suited for.

Chart Type	Best For	Key Insights Provided	Data Nature	Limitations
Bar Chart	- Comparing discrete categories/groups.	- Relative magnitudes (e.g., sales by product).	- **Categorical** (nominal/ordinal) data.	- Cluttered with too many categories.
	- Showing rankings or distributions.	- Outliers or dominance in categories.	- Counts, frequencies, or aggregated values.	- Less effective for trends over time.
Line Chart	- Tracking trends over continuous intervals.	- Trends, growth, or decline (e.g., stock prices).	- **Time-series** or sequential data.	- Misleading if time intervals are uneven.
	- Comparing multiple series over time.	- Seasonality, cycles, or correlations.	- Continuous numerical variables.	- Overplotting with too many lines.
Pie Chart	- Showing parts of a whole (composition).	- Proportional contributions (e.g., market share).	- **Percentages** summing to 100%.	- Hard to compare small slices.
	- Highlighting dominant categories.	- Dominant vs. minor components.	- Few (ideally <6) categories.	- Ineffective for precise comparisons.

1. **Bar Charts:**
    - **Use when:** Comparing distinct categories
    - **Avoid when:** Displaying trends over time (line charts are better).
2. **Line Charts:**
    - **Use when:** Showing changes over time
    - **Avoid when:** Data is categorical or lacks a sequential order.
3. **Pie Charts:**
    - **Use when:** Emphasizing proportions of a whole

- **Avoid when:** Categories are numerous or values are similar (bar charts are clearer).