**UPTRAX TECHNOLOGIES LIMITED**

**WEEKLY PROJECTS FOR BEGINNERS**

You've been given resources to aid in responding to these queries. This week is dedicated to an introduction to Data Analytics and basic analysis of a Customer Sales Dataset, aiming to enhance your comprehension of fundamental terms utilized in Data Analysis. Please ensure you review the materials before tackling the questions. Best of luck!

**Questions for Introduction to Data Analysis.**

1a. Explain Data Analytics in broader terms.

Data analytics refers to the whole process of data management: data collecting, storing, organizing, and analysing. It includes the tools and techniques used to deep-dive into data, as well as those used to communicate the results.

1b. Explain how data analytics can be used to gain good market insights.

Data analytics is very good for market research. It’s helps to identify new targets markets. Researchers examine psychographic data as well demographic and behavioural data. Business use data to identify target audience that tends to engage their products and services. It helps the organisation identify strategies for success.

2a. Distinguish between a Data Analyst, Data Scientist and Data Engineer.

Data Analyst work on structured data to solve business problems using data tool like SQL, Python, R, Programming languages, statistical analysis and data visualization software. They acquire data from primary and secondary sources.

Data Scientist work on a broader scope. They work with the unknown by using more advanced data techniques to make predictions about the future. They create their own modelling process to handle both structured and unstructured data. They gather, clean and processing raw data.

Data Engineer builds systems that collect, manage and convert raw data into usable information for data scientist and analysts to interpret. They make data accessible so that organisation can use it to evaluate and optimize their performance.

2b. Briefly explain how data analytics can be used in sports, healthcare and finance sectors.

Sport sector

Analyst study the athletic performance and business health to optimize the processes and the success of a sport organization. It helps to make important strategic decisions on and off the field. Decision helps to improve in game strategies, nutrition plans, other areas to boost athletes’ performance. Decisions helps to business side of the sport like ticket sales, merchandise sale, fan engagements and many others. They tend to increase profitability.

Healthcare Sector

With the use of historic and current data, healthcare sector is provided with actionable insights to improve decision making and optimize outcomes in their industry. It also helps to improve the patient experience and health outcomes. Data analysis in the health sector offers the potential to reduce operation cost, improve efficiency, and treat patients.

Finance Sector

Data analysis in financial sector helps to reduce component of human error from daily financial transactions. It helps finance executives to turn structured or unstructured data into insights that promote better decision making. It helps finance teams gather a clear view of key performance indicators such as revenue, generated, net income, payroll etc.

3a. Provide an example of a situation where predictive analytics could be applied for better decision-making,

Predictive analysis uses historical data to make inferences about future events.

Example

A retailing firm would use predictive analytics to determine which products are most likely purchased together and offer discounts on those items. They can use data to identify customers at risks of leaving for a competitor.

3b. Provide an example of a business scenario where descriptive analytics would be valuable.

Descriptive statistics data analysis describes, summarizes, and visualizes the basic features of data, through charts and reports.

Examples

A clothing store can use the descriptive analytics to identify the year on year percentage sales growth, revenue per customer and profit from various items.

4a. List six formats of data and explain three.

1. Character format

This describes text attributes which are applied at character level. E.g font and size formatting

1. Numeric data type

These are types of data that consist of numbers, which can be computed mathematically with various standard operators such as add, minus, multiply, divide and more.

1. Graphic format

Data is stored in points, pixels or grids, and the format that work with vectors.

1. Date Data Type
2. Basing Pointer Data Type
3. UCS-2 Format

4b. What distinguishes primary data from secondary data in terms of collection and usage.

|  |  |  |
| --- | --- | --- |
|  | Primary data | Secondary data |
| Collection | Original and first hand. Personal experience or evidence. Raw data. | Interpretation and analysis of the primary data. Already collected by someone. |
| Usage | No precaution to use the data. Collected with a definite purpose | Precaution and editing are essential as data was collected by another person. |

4c. How does quantitative and qualitative data differ in terms of analysis and interpretation?

|  |  |  |
| --- | --- | --- |
|  | Quantitative data | Qualitative data |
| analysis | Can be analysed statistically and calculated into averages, means, and other numerical data points | Represents opinions or feelings and cannot be represented by a numerical statistic  It involves a more complex system |
| interpretation | It used to answer precise question and prove or disprove hypotheses | Provide richer insights on a smaller scale. |

5a. Explain big data, small data and metadata in your own terms.

1. Big data

The large pictures that encompass many different types of data and its mainly unstructured

1. Small data

The small picture, it is structured, focused, and easily interpreted.

1. Meta data

It provides information about one or more aspects of data.

5b. List and explain the 8 steps of data lifecycle.

1. Generation: for data to begin, the data must first be generated.
2. Collection: not all data collected is used. Its up to the data team to identify what information should be captured and best means for doing so.
3. Processing: once data has been collected, it must be processed
4. Storage: after data has been collected and processed, it must be stored for future use
5. Management: data management involves organizing, storing, and retrieving data as necessary over life of a data project
6. Analysis: this is the process that attempt to glean meaningful insights from raw data.

6a. Briefly explain how statistics is used in data analytics.

The two main statistical method used in data analysis are descriptive statistics which summarizes data using indexes such as mean and median and another inferential statistic which draw conclusion from data using statistical test.

6b. What is Data mining?

Data mining aka knowledge discovery in data (KDD) is the process of detecting anomalies, patterns, and relationships to predict outcomes.

Context: Applications of data mining include predicting how markets will behave, allowing businesses to [anticipate customer needs](https://monkeylearn.com/blog/customer-needs-analysis/) and be proactive

7a. List 2 tools that can be used for data visualization.

1. Tableau
2. Power BI

7b. List 2 programming languages that can be used for data analysis.

1. SQL
2. JAVA

8a. List and Explain 3 python libraries.

1. NumPy: this offers comprehensive mathematical function, random numbers generator, linear algebra routines, fourier transforms and more.
2. Pandas: is a powerhouse data manipulation and analysis, offering powerful, expressive, flexible data structures. It allows fast data cleaning, preparation and analysis.
3. Matplotlib: is a comprehensive library for creating static, animated, and interactive visualization in python.

8b. What are some of the challenges associated with processing and analyzing unstructured data compared to structured data?

One of the main challenges associated with working with unstructured data is that it can be difficult to organize and analyse unlike structured data which is organized into predetermined format, unstructured data can come in many forms, such as text documents, images and videos.

**Questions for Customers Sale Dataset Analysis**

1. Basic Information:

- What is the total number of invoices in the dataset?

99457

- How many unique customers are there?

More than 10,000 (99457)

- What is the average age of customers?

43.42709

2. Category Analysis:

- What are the unique categories of items in the dataset?

Books, Clothing, Cosmetics, Food & Beverage, Shoes, Souvenir, Technology, Toys

- Which category has the highest quantity sold?

Clothing 103,558

- Can you find the average price for each category?

|  |  |
| --- | --- |
| **Row Labels** | **Average of price** |
| Books | 46 |
| Clothing | 901 |
| Cosmetics | 122 |
| Food & Beverage | 16 |
| Shoes | 1807 |
| Souvenir | 35 |
| Technology | 3157 |
| Toys | 108 |
| **Grand Total** | **689** |

3. Payment Method Distribution:

- What is the different payment methods used?

Cash, Credit Card, Debit Card

- Create a pie chart to represent the distribution of payment methods.

4. Shopping Mall Insights:

- Which shopping mall had the most transactions?

Mall of Istanbul, 19,943 transactions

- What is the average quantity of items purchased in each shopping mall?

|  |  |
| --- | --- |
| **Row Labels** | **Average of quantity** |
| Cevahir AVM | 2.995 |
| Emaar Square Mall | 3.014 |
| Forum Istanbul | 3.002 |
| Istinye Park | 3.012 |
| Kanyon | 2.999 |
| Mall of Istanbul | 3.014 |
| Metrocity | 2.991 |
| Metropol AVM | 3.005 |
| Viaport Outlet | 2.995 |
| Zorlu Center | 3.002 |
| **Grand Total** | **3.003** |

5. Temporal Analysis:

- How many invoices are there for each month?

|  |  |
| --- | --- |
| **Row Labels** | **Count of invoice\_no** |
| **Jan** | **11608** |
| **Feb** | **10482** |
| **Mar** | **8730** |
| **Apr** | **7487** |
| **May** | **7697** |
| **Jun** | **7581** |
| **Jul** | **7877** |
| **Aug** | **7635** |
| **Sep** | **7353** |
| **Oct** | **7764** |
| **Nov** | **7563** |
| **Dec** | **7680** |
| **Grand Total** | **99457** |

- Create a line chart to visualize the trend in sales over time.

6. Customer Gender Distribution:

- How many male and female customers are in the dataset?

|  |  |
| --- | --- |
| **Row Labels** | **Count of customer\_id** |
| Female | 59479 |
| Male | 39974 |
| **Grand Total** | **99453** |

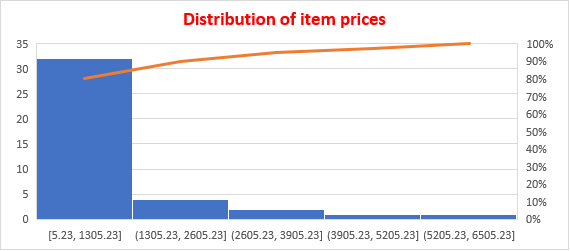
- Create a bar chart to represent the gender distribution.

7. Price Analysis:

- What is the total revenue generated from all invoices?

|  |  |
| --- | --- |
| **Row Labels** | **Total Revenue** |
| **Both** | **$ 2,198** |
| **Female** | **$40,930,203** |
| **Male** | **$27,618,964** |
| **Grand Total** | **$68,551,366** |

- Create a histogram to show the distribution of item prices.



8. Age Group Analysis:

- Group customers into age brackets (e.g., 20-30, 31-40) and count how many customers are in each group.

|  |  |
| --- | --- |
| **Row Labels** | **Count of customer\_id** |
| 11-20 | 5624 |
| 21-30 | 19400 |
| 31-40 | 19266 |
| 41-50 | 19066 |
| 51-60 | 18932 |
| 61-70 | 17169 |
| **Grand Total** | **99457** |

- Create a bar chart to represent the distribution of customers across age groups.

9. Correlation Analysis:

- Is there any correlation between age and the quantity of items purchased?

- Create a scatter plot to visualize any potential relationships.

10. Top Customers:

- Identify the top 5 customers based on the total amount spent.

|  |  |
| --- | --- |
| **Price** | **Count of customer\_id** |
| $ 2,401 | 2022 |
| $ 3,001 | 2013 |
| $ 3,150 | 1009 |
| $ 4,200 | 1001 |
| $ 5,250 | 1001 |
| **Grand Total** | **7046** |

- Create a bar chart to represent the spending of these top customers.

**SUBMISSION DATE: FRIDAY 8AM - SATURDAY 8AM**