# Cover Page

Data and Web Analytics Coursework

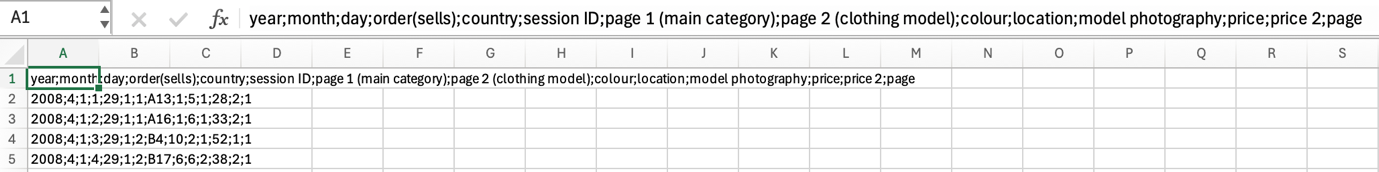
Phạm Thanh Trúc

23/07/2024

**Table of Contents**

## Task a:

Original excel file



2 ways to separate data:

1. in Excel

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A table of numbers and letters

Description automatically generated

1. by coding in R

A screenshot of a computer program

Description automatically generated

Columns were separated

A table of numbers with text

Description automatically generated with medium confidence

Extracted Columns:

1. **Month**: Helps identify seasonal trends and monthly variations in sales, revealing if sales peak during specific months, such as during summer sales or back-to-school shopping periods.
2. **Day**: The specific day of the month can be used to analyze daily trends, identify patterns related to weekdays vs. weekends, and pinpoint specific high or low sales days.
3. **Order(sells)/Click\_stream**: Represents the sequence of clicks during one session, which is crucial for understanding customer behavior, click paths, and navigation patterns within the e-store.
4. **Country**: Identifies the country of origin of the IP address, providing insights into the geographical distribution of customers, which can influence marketing strategies, inventory distribution, and localization of the e-store.
5. **Session ID**: This unique identifier for each session allows tracking of individual user sessions, essential for session analysis, understanding user behavior, and measuring engagement levels.
6. **Page 1 (Main Category)**: Knowing the main product category that customers are interested in helps analyze product category performance, identify popular categories, and make inventory decisions.
7. **Colour**: Analyzing color preference helps understand customer preferences, trends, and aids in inventory planning for popular colors.
8. **Location**: The photo location on the page impacts customer engagement and product visibility, useful for optimizing website layout and product placement.
9. **Model Photography**: Indicates whether the product is shown in a front view or profile, which can affect customer perception and engagement. Analyzing this can help in optimizing product photography for better sales.
10. **Price**: Fundamental for revenue analysis, pricing strategy evaluation, and understanding the impact of price changes on sales.
11. **Price 2**: Indicates whether the price is higher than the average price for the product category, useful for pricing strategy analysis and understanding the market positioning of products.
12. **Page**: The page number within the e-store helps analyze navigation patterns, customer engagement across different pages, and optimize the e-store structure for better user experience.
13. **Revenue**: This is a key metric for understanding the financial performance of the e-store. It helps analyze the total sales generated, identify high-revenue products, and assess the overall profitability of the business.

Correlation matrix

A graph with red and blue squares

Description automatically generated

Explain

## Task b:

Sales by Month Chart

A graph of blue bars

Description automatically generated

The bar chart titled "Total product\_sold by Month" shows the number of products sold during each month from April (4) to August (8). The chart reveals the following:

* **April (4)**:
  + Sales are approximately 6,900 units.
  + This is the highest sales month in the dataset, reflecting strong market demand or successful sales strategies.
* **May (5)**:
  + Sales decrease to roughly 5,200 units.
  + This is a decrease of 1,700 units from April, a 24.6% decline.
* **June (6)**:
  + Sales slightly decrease further to approximately 4,800 units.
  + This is a decrease of 400 units from May, a roughly 7.7% decline.
* **July (7)**:
  + Sales increased slightly to approximately 5,100 units.
  + This is an increase of 300 units from June, a roughly 6.25% increase.
* **August (8)**:
  + Sales drop significantly to approximately 1,900 units.
  + However, this drop is due to data being collected only until the 13th day of August.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

### Summary for Sales Chart:

* The highest sales occur in April, followed by a sharp decline in May.
* Sales remain relatively stable from May to July.
* The significant drop in sales observed in August is primarily due to incomplete data, not necessarily a true reflection of market demand.

### Possible Reasons:

1. **Seasonal Demand**: April might have coincided with the launch of new products, a spring sale, or other promotional events, leading to a surge in sales. After April, the lack of such events or the end of a season could have led to the drop in sales. The relatively stable sales from May to July suggest consistent demand.
2. **Marketing and Promotions**: April's high sales could be a result of effective marketing campaigns or promotions that were not sustained in the following months.
3. **Inventory and Product Availability**: The decline after April might also suggest that popular products were out of stock or that the product lineup did not attract as much interest.
4. **Customer Behavior**: There could be a pattern in customer behavior where interest wanes as the summer progresses, possibly due to vacations or shifts in spending habits.
5. **Grand opening event**: The significantly high number of sold units on the first and second days in April, nearly double the sales of subsequent days, could be attributed to a grand opening event. This initial spike in sales is often seen during such promotions, as customers are drawn in by special offers and the excitement surrounding the launch.

A graph of blue bars

Description automatically generated

A graph with text and numbers

Description automatically generated with medium confidence

A graph with numbers and lines

Description automatically generated

A graph with blue bars

Description automatically generated

A graph of blue rectangular bars

Description automatically generated with medium confidence

A graph of blue bars

Description automatically generated

A graph with blue squares

Description automatically generated

A graph with blue squares

Description automatically generated

A graph with blue squares

Description automatically generated

## Task c:

A close-up of a text

Description automatically generated

A table of numbers with a number on it

Description automatically generated

Revenue by Month Chart

A graph of blue bars

Description automatically generated

Analysis

* **April (4)**:
  + Revenue is approximately 300,000 US dollars.
  + This is the highest revenue month in the dataset, indicating strong performance.
* **May (5)**:
  + Revenue drops to approximately 220,000.
  + This is a decrease of 80,000 from April, a roughly 26.7% decline.
* **June (6)**:
  + Revenue continues to decline, reaching approximately 210,000.
  + This is a decrease of 10,000 from May, a roughly 4.5% decline.
* **July (7)**:
  + Revenue slightly increases to approximately 220,000.
  + This is an increase of 10,000 from June, a roughly 4.8% increase.
* **August (8)**:
  + Revenue drops significantly to approximately 180,000.
  + However, this drop is due to data being collected only until the 13th day of August.

### Summary for Revenue Chart:

* The highest revenue occur in April, followed by a sharp decline in May.
* Revenue remain relatively stable from May to July.
* The significant drop in revenue observed in August is primarily due to incomplete data, not necessarily a true reflection of market demand.

### Possible Reasons:

1. **Seasonal Demand**: April might have coincided with the launch of new products, a spring sale, or other promotional events, leading to a surge in revenue. After April, the lack of such events or the end of a season could have led to the drop in revenue. The relatively stable revenue from May to July suggest consistent demand.
2. **Marketing and Promotions**: April's high revenue could be a result of effective marketing campaigns or promotions that were not sustained in the following months.
3. **Inventory and Product Availability**: The decline after April might also suggest that popular products were out of stock or that the product lineup did not attract as much interest.
4. **Customer Behavior**: There could be a pattern in customer behavior where interest wanes as the summer progresses, possibly due to vacations or shifts in spending habits.
5. **Grand opening event**: The significantly high number of revenue on the first and second days in April, nearly double the revenue of subsequent days, could be attributed to a grand opening event. This initial spike in sales is often seen during such promotions, as customers are drawn in by special offers and the excitement surrounding the launch.

**April**

* **Strong Performance**: April had the highest revenue, possibly due to seasonal demand or successful marketing strategies that attracted more customers.
* **Promotional Campaigns**: Special promotions or events in April might have driven higher sales, contributing to the revenue peak.
* **Grand Opening Event:** The unusually high revenue recorded on the first and second days of April, nearly double that of the following days, may be linked to a grand opening event. Such promotional events typically result in an initial surge in sales, as customers are enticed by special deals and the excitement of the launch.

Revenue by Day in April Chart

A graph of blue bars

Description automatically generated

**May**

* **Post-Peak Adjustment**: The drop in revenue in May (approximately 26.7%) may indicate a natural adjustment after the peak month of April, as customers may have exhausted their purchasing during the promotional period.
* **Increased Competition**: The entrance of competitors or alternative products may have drawn customers away, impacting sales.
* **Consumer Behavior**: Changes in consumer spending habits or seasonal trends may also contribute to the decline.

**June**

* **Continued Decline**: The slight decrease in June (approximately 4.5%) could suggest that the initial drop in May continued, potentially due to lingering effects of increased competition or reduced consumer interest.
* **Market Saturation**: By June, the market may have reached saturation, leading to less urgency in purchases from consumers.

**July**

* **Minor Recovery**: The small increase in July (approximately 4.8%) could indicate a rebound as marketing efforts were adjusted or new products were introduced.
* **Seasonal Trends**: July may benefit from summer sales or promotions, attracting more customers and encouraging spending.

**August**

* **Limited Data**: The significant drop in August (to approximately 180,000) should be viewed with caution, as it only reflects data collected until the 13th day.
* **Seasonal Lull**: If data collection continued for the full month, revenue may have picked up again as seasonal promotions began or as consumers prepared for back-to-school shopping.

Total product sold by category chart

A graph of blue rectangular bars

Description automatically generated with medium confidence

Total revenue by category chart

A graph of blue bars

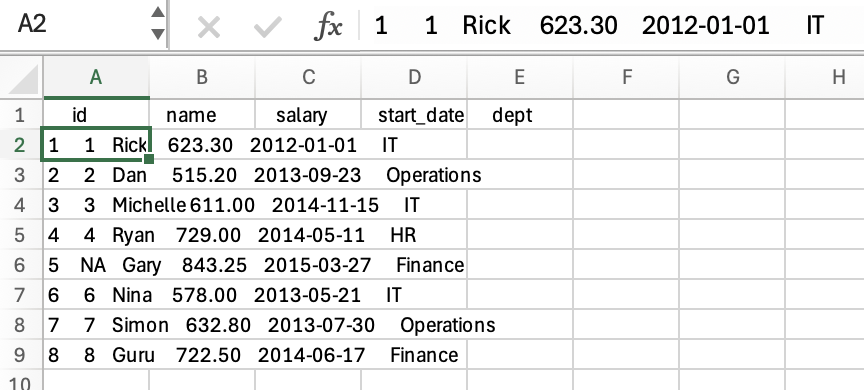
Description automatically generated with medium confidence

Insight on the 2 above charts:

|  |  |
| --- | --- |
| **Product Category** | **Recommendations** |
| **Trousers (Category 1)** | Because many trousers were sold and brought good revenue so we just need to:  - Maintain quality and availability to sustain this performance.  - Introduce premium or specialized products to further boost revenue.  - Cross-sell with other categories to increase sales in lower-performing categories. |
| **Skirts (Category 2)** | Because skirts were sold less than trousers but brought as high revenue as trousers so we should **increase selling skirts** by:  - Promote through special collections or limited editions to capitalize on their strong revenue potential.  - Expand the range with new styles to capture more market share.  - Implement targeted marketing campaigns to increase visibility and sales. |
| **Blouses (Category 3)** | - Although many blouses were sold but its revenue is not as much as it should be so we should **increase price of blouses** |
| **Sale Items (Category 4)** | Because sale brought small profit so we should:  **- reduce the number of items on sale**,  - focusing on moving only slow-moving inventory to improve overall profitability.  - Offer bundling strategies to clear inventory.  - Run limited-time promotional events to drive urgency.  - Review inventory to phase out outdated or unpopular items. |

## Task d:

Original excel file



Data after being cleaned

A screenshot of a table

Description automatically generated

Descriptive Statistics for Salary

A screenshot of a computer code

Description automatically generated

A white background with black text

Description automatically generated

### Summary:

* The mean salary is around 656.88, indicating the average salary among the employees.
* The median salary is 628.05, showing the middle point of the salary distribution.
* There is no single mode, as all salaries are unique.
* The standard deviation is approximately 103.06, indicating the dispersion of salaries around the mean.
* The variance is 10621.25, representing the spread of the salary data.

## Task e:

1. Identify actors with eye colors other than black and a height exceeding 150 cm

A screenshot of a computer program

Description automatically generated

A table of names and numbers

Description automatically generated

1. The procedure for adding a Body Mass Index (BMI) column.

A white background with black text

Description automatically generated

A table of names and numbers

Description automatically generated

Plot of height versus BMI

A screenshot of a white screen

Description automatically generated

A black text on a white background

Description automatically generated

A close-up of words

Description automatically generated

A table of names and colors

Description automatically generated

A grid of white lines

Description automatically generated

The graph plots Height (in centimeters) on the x-axis against BMI (Body Mass Index) on the y-axis for characters from the Star Wars dataset. Here’s a breakdown of what the graph shows:

1. **Overall Trend**: Most data points are clustered at the bottom of the graph, indicating that the BMI values for the majority of the characters are relatively low. The height of these characters ranges between approximately 150 cm and 220 cm.
2. **Outlier**: There is one significant outlier in the graph with a BMI much higher than the others. This character has a BMI of over 400, which is extremely high and unusual compared to the rest of the dataset.
3. **Height Distribution**: The height of the characters varies between around 65 cm and 250 cm, but most characters have heights between 150 cm and 200 cm.
4. **BMI Distribution**: Except for the outlier, the BMI values are mostly clustered below 100. This suggests that, for the majority of characters, their weight and height result in a lower BMI, which is consistent with normal to overweight categories in humans.

### Possible Interpretation:

* The outlier could represent a character with an unusual physique or an error in the data entry, as such a high BMI is not typical.
* The rest of the characters follow a more expected distribution, where the BMI increases slightly with height but remains within a typical range.
* Characters that are shorter generally have a slightly lower BMI, which might indicate that they are lighter relative to their height.

The dataset seems to have a variety of characters with different body types, but the extreme outlier might require further investigation to understand if it’s a valid data point or an anomaly.

A screenshot of a computer

Description automatically generated



The outlier in the graph represents the character **Jabba Desilijic Tiure**, commonly known as Jabba the Hutt, from the Star Wars universe.

* + **Height**: 175 cm
  + **Mass**: 1358.0 kg
  + **BMI**: 443.42857

He has a body type that is extremely different from the more humanoid characters in the dataset. This leads to an exceptionally high BMI, which is reflected in the graph as an extreme outlier.

Jabba’s data point appears far above the rest, showing that his BMI is an extreme anomaly. The majority of the characters have BMIs that are under 60), making Jabba’s BMI of over 440 stand out significantly.

This also reflects how the BMI formula, designed for humans, might not accurately represent the body types of such non-humanoid creatures.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A graph showing a number of points

Description automatically generated