NATIONAL INSTITUTE OF BUSINESS MANAGEMENT

School of Computing

Higher National Diploma in Software Engineering

Batch – GAHDSE241F

Data Warehousing and Business Intelligence - Coursework

Module Lecturer: Niranga Dharmaratna

Name	C.D. Wijesekara
Index #	GAHDSE241F - 045
Submission:	2024.12.05

Ethical Declaration of Original Work

I declare that the work presented in this coursework is entirely my own. I confirm that:

1. The work presented in this coursework is conducted by me, and any contributions from other individuals are appropriately acknowledged.
2. Any external sources of information and ideas used in this work are cited and referenced accurately. I have provided proper credit to the original authors through citations in the text and a comprehensive list of references.
3. The data and findings presented in this work are genuine and have not been manipulated or fabricated. Any assistance received in the collection and analysis of data is acknowledged appropriately.
4. I have not submitted this work, or any part of it, for any other academic qualification.
5. I understand the ethical principles governing academic work, including honesty, integrity, and accountability. I have adhered to these principles throughout the process.

I am aware of the consequences of academic misconduct and understand that any violation of ethical standards may result in disciplinary action.

Signed: Chithmi
Chithmi Dilaksha Wijesekara
2024.12.0

Table of Contents

Table of Figures

1.0 INTRODUCTION

The goal of this project is to build a Data Warehouse (DW) using Google Big Query and analyze the data using Tableau to efficiently analyze Fashion Dataset UK-US. Leveraging ADW's cloud-based capabilities, the dataset is stored, processed, and structured for seamless querying and analysis. Tableau, a leading visualization tool, accomplishes this by transforming raw data into interactive dashboards and reports. This dataset is a valuable resource for researchers, industry professionals, and analysts, providing actionable insights for informed decision-making in the highly dynamic fashion industry.

The Fashion Sales Dataset is a comprehensive resource that provides in-depth insights into the sales processes of the fashion industry. Designed to simulate real-world sales scenarios, this dataset is a valuable tool for analyzing sales trends and strategizing for success in the fashion market.

Here Included Key Features:

1. Realistic Sales Data – Customer purchase and transaction details include product-specific attributes such as names, prices, brands, types, and descriptions.
2. Different Product Attributes – Data fields such as ratings, review counts, available sizes, colors, and purchase history. Enables analysis of customer preferences and product performance.
3. Simulated Customer Interactions - Integrates insights from fashion magazines, influencers, customer reviews, and social media comments.
4. Seasonal and Time-Based Analysis - Captures data across different seasons and specific time periods, providing insights into seasonal preferences and trends.

2.0 PREREQUISITES

2.1 Required Software and Tools

Sample Dataset - Kaggle

- A large publicly available dataset for data storage and visualization (minimum size: 100MB). Example sources include.
- Tableau Public Sample Datasets.

Google Search Console

- To Find Free Dataset for analysis

Google Colab

- Clean Dataset – Remove Duplicate Records, Multiple values and etc.
- Formalize the data set as required for this analysis.

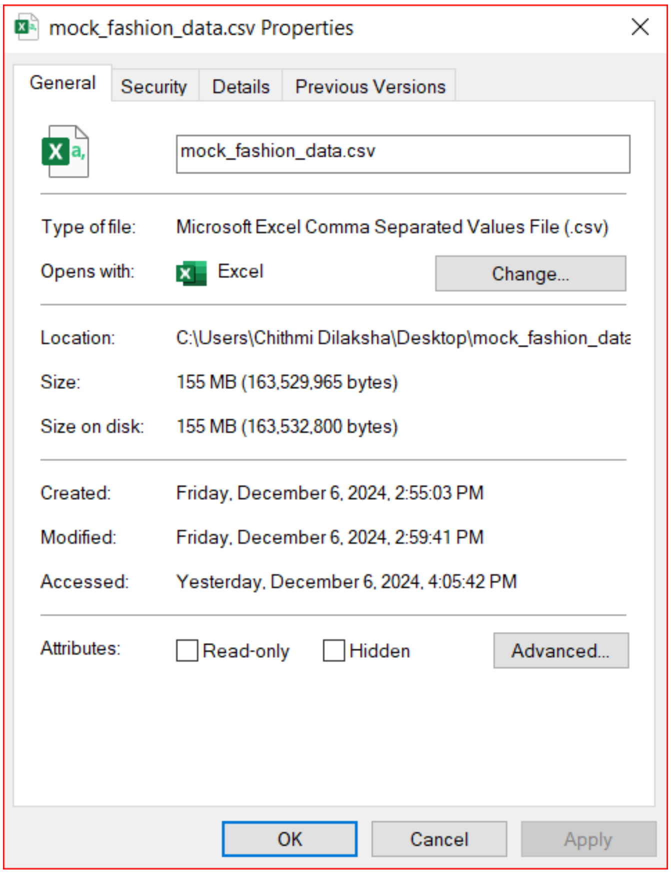
Goggle Big Query

- Create data warehouse using Google Big Query
- Create Dataset and Query

3.0 COURSEWORK TASKS

3.1 Task 1: Sourcing and Preparing Sample Data

1. Select Dataset
   Here, a dataset of over 100MB was selected.



mock_fashion_data.csv

Type of file: Microsoft Excel Comma Separated Values File (.csv)

Opens with: Excel

Location: C:\Users\Chithmi Dilaksha\Desktop\mock_fashion_data

Size: 155 MB (163,529,965 bytes)

Size on disk: 155 MB (163,532,800 bytes)

Created: Friday, December 6, 2024, 2:55:03 PM

Modified: Friday, December 6, 2024, 2:59:41 PM

Accessed: Yesterday, December 6, 2024, 4:05:42 PM

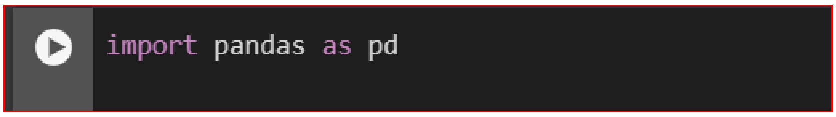
Attributes: □ Read-only □ Hidden

Figure 01- Uncleared data set

2. Clean Data Set

This set contains a large amount of data and **Google Colab** panda library was used to efficiently process the data and maintain uniformity by removing empty values, outliers, and duplicate values.

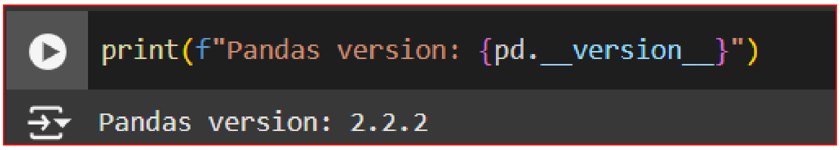
- Step 01



```
import pandas as pd
```

Figure 02 - Import Pandas version

- Step 02

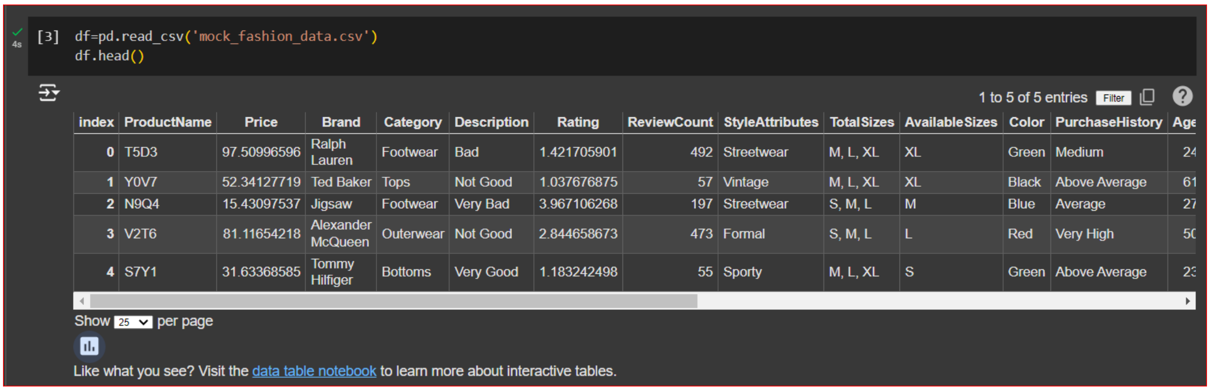


```
print(f"Pandas version: {pd.__version__}")
```

```
Pandas version: 2.2.2
```

Figure 03 - Display Pandas Version

- Step 03



```
[3] df=pd.read_csv('mock_fashion_data.csv')
    df.head()
```

1 to 5 of 5 entries

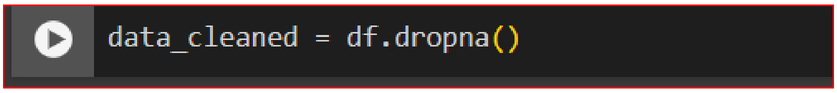
index	ProductName	Price	Brand	Category	Description	Rating	ReviewCount	StyleAttributes	TotalSizes	AvailableSizes	Color	PurchaseHistory	Age
0	T5D3	97.50996596	Ralph Lauren	Footwear	Bad	1.421705901	492	Streetwear	M, L, XL	XL	Green	Medium	24
1	Y0V7	52.34127719	Ted Baker	Tops	Not Good	1.037676875	57	Vintage	M, L, XL	XL	Black	Above Average	61
2	N9Q4	15.43097537	Jigsaw	Footwear	Very Bad	3.967106268	197	Streetwear	S, M, L	M	Blue	Average	27
3	V2T6	81.11654218	Alexander McQueen	Outerwear	Not Good	2.844658673	473	Formal	S, M, L	L	Red	Very High	50
4	S7Y1	31.63368585	Tommy Hilfiger	Bottoms	Very Good	1.183242498	55	Sporty	M, L, XL	S	Green	Above Average	23

Show 25 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

Figure 04 - Read Unclear Dataset Data

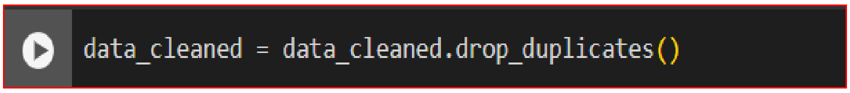
- Step 04



```
data_cleaned = df.dropna()
```

Figure 05 - Drops Rows with Missing (Non) Values

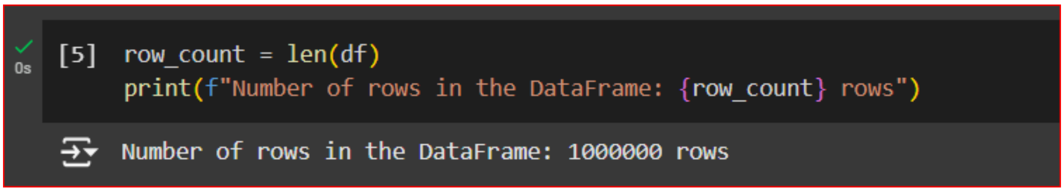
- Step 05



```
data_cleaned = data_cleaned.drop_duplicates()
```

Figure 06 - Remove Duplicate Values

- Step 06

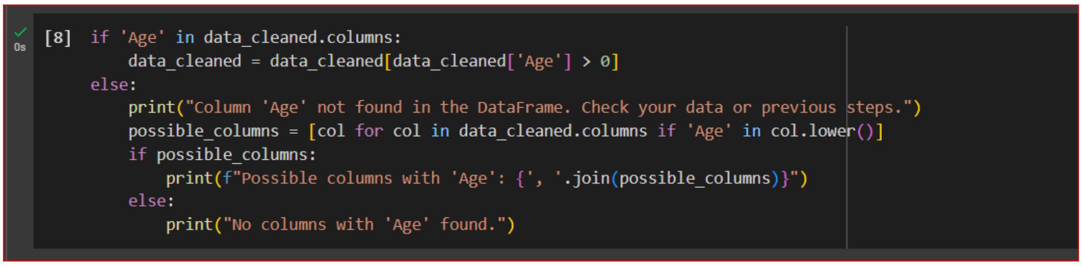


```
row_count = len(df)
print(f"Number of rows in the DataFrame: {row_count} rows")
```

```
Number of rows in the DataFrame: 1000000 rows
```

Figure 07 - Display Row Count Before Clean

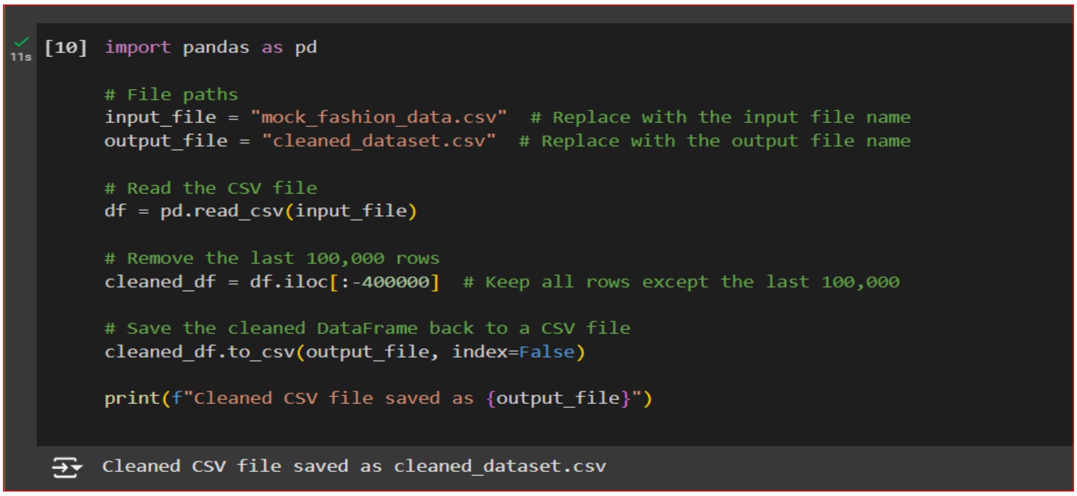
- Step 07



```
if 'Age' in data_cleaned.columns:
    data_cleaned = data_cleaned[data_cleaned['Age'] > 0]
else:
    print("Column 'Age' not found in the DataFrame. Check your data or previous steps.")
    possible_columns = [col for col in data_cleaned.columns if 'Age' in col.lower()]
    if possible_columns:
        print(f"Possible columns with 'Age': {', '.join(possible_columns)}")
    else:
        print("No columns with 'Age' found.")
```

Figure 08 - Clean Data Column

- Step 08

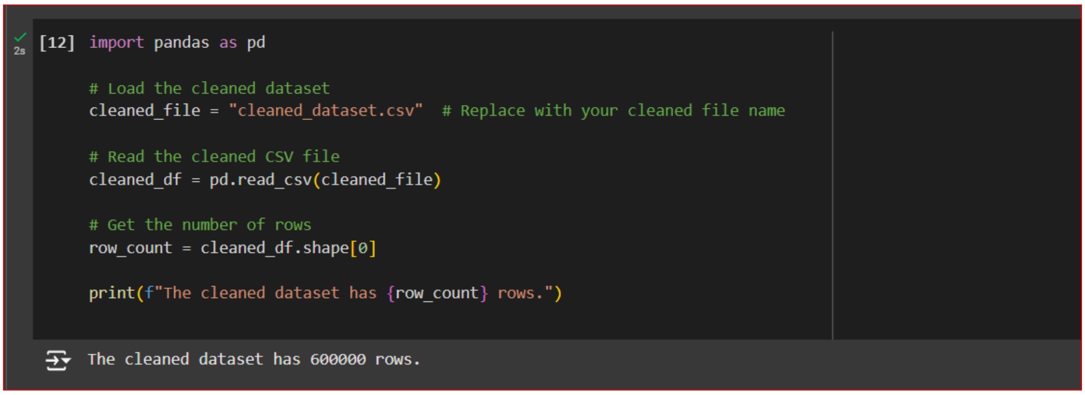


```
import pandas as pd

# File paths
input_file = "mock_fashion_data.csv"  # Replace with the input file name
output_file = "cleaned_dataset.csv"  # Replace with the output file name

# Read the CSV file
df = pd.read_csv(input_file)

# Remove the last 100,000 rows
cleaned_df = df.iloc[:-400000]  # Keep all rows except the last 100,000

# Save the cleaned DataFrame back to a CSV file
cleaned_df.to_csv(output_file, index=False)

print(f"Cleaned CSV file saved as {output_file}")
```

```
Cleaned CSV file saved as cleaned_dataset.csv
```

Figure 09 - Remove Rows for Reduce the File Size

- Step 09

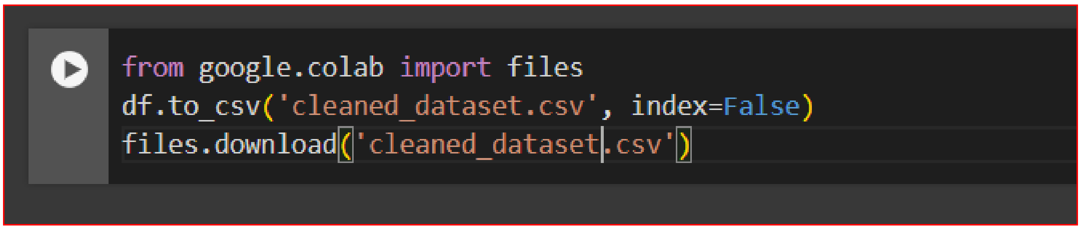


```
import pandas as pd

# Load the cleaned dataset
cleaned_file = "cleaned_dataset.csv"  # Replace with your cleaned file name

# Read the cleaned CSV file
cleaned_df = pd.read_csv(cleaned_file)

# Get the number of rows
row_count = cleaned_df.shape[0]

print(f"The cleaned dataset has {row_count} rows.")
```

```
The cleaned dataset has 600000 rows.
```

Figure 10 - Display Last Modified Data Set

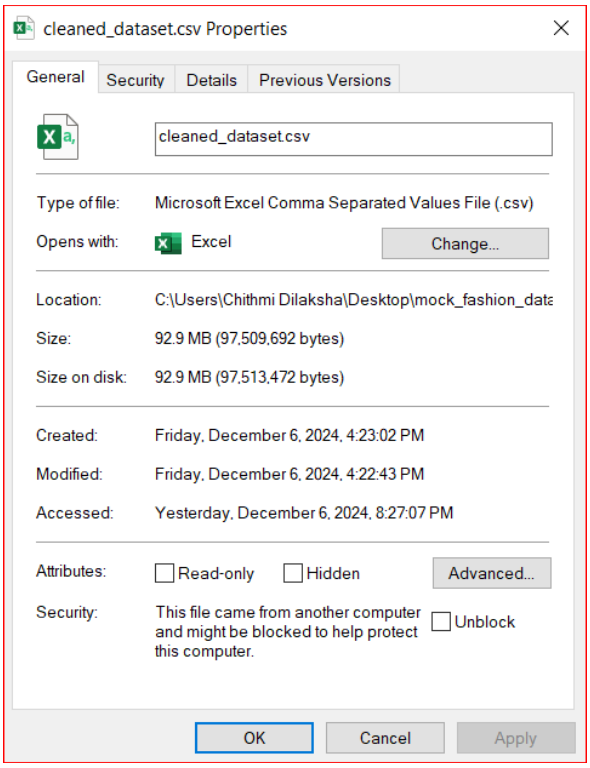
- Step 10



```
from google.colab import files
df.to_csv('cleaned_dataset.csv', index=False)
files.download('cleaned_dataset.csv')
```

Figure 11 - Download Clean Dataset

3. After Cleaned Data Set



cleaned_dataset.csv

Type of file: Microsoft Excel Comma Separated Values File (.csv)

Opens with: Excel

Location: C:\Users\Chithmi Dilaksha\Desktop\mock_fashion_data

Size: 92.9 MB (97,509,692 bytes)

Size on disk: 92.9 MB (97,513,472 bytes)

Created: Friday, December 6, 2024, 4:23:02 PM

Modified: Friday, December 6, 2024, 4:22:43 PM

Accessed: Yesterday, December 6, 2024, 8:27:07 PM

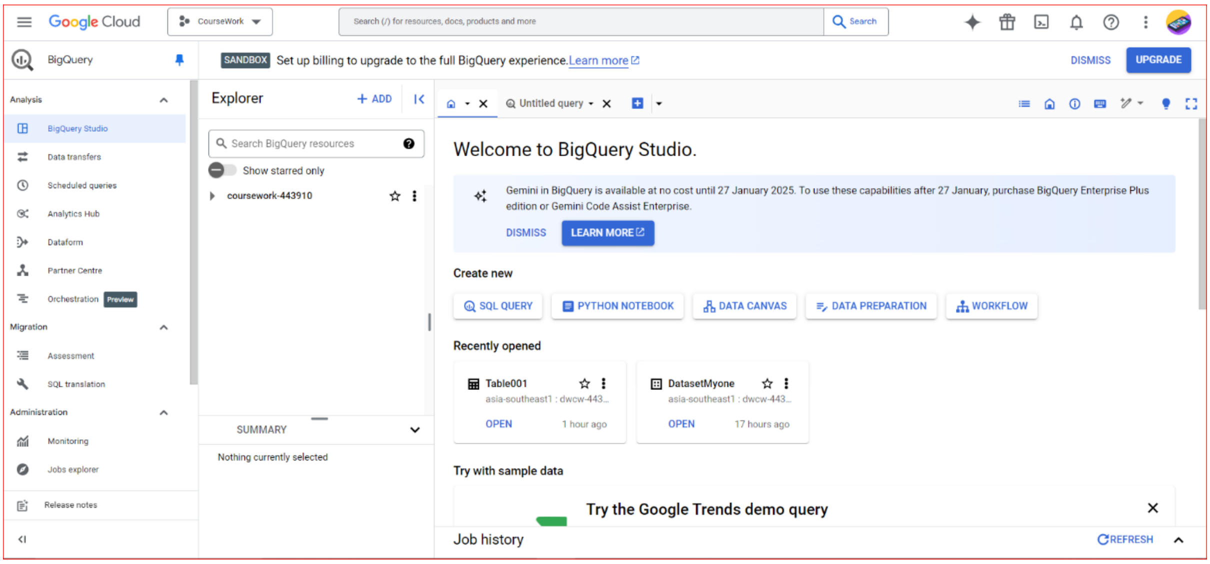
Attributes: □ Read-only □ Hidden

Security: This file came from another computer and might be blocked to help protect this computer. □ Unblock

Figure 12 - Clean Dataset Size

3.2 Task 2: Amazon Redshift Account Create for Cloud Data Warehousing

- Step 01 – Create New Project



Set up billing to upgrade to the full BigQuery experience. Learn more

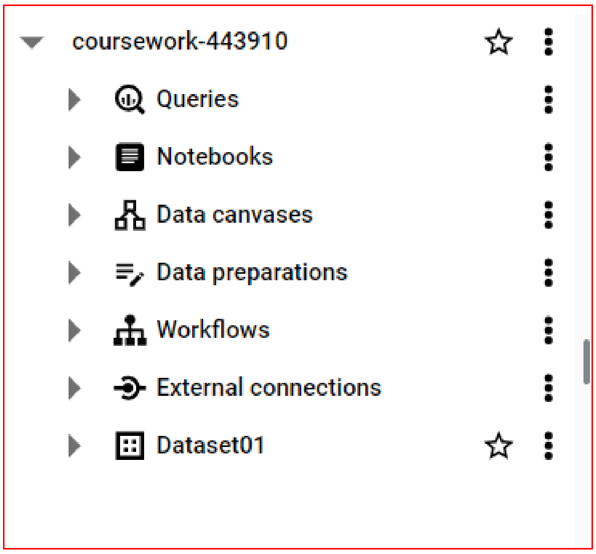
Welcome to BigQuery Studio.

Gemini in BigQuery is available at no cost until 27 January 2025. To use these capabilities after 27 January, purchase BigQuery Enterprise Plus edition or Gemini Code Assist Enterprise.

Try the Google Trends demo query

Figure 13 - Create New Project

- Step 02 – Create Data Set



coursework-443910
- Queries
- Notebooks
- Data canvases
- Data preparations
- Workflows
- External connections
- Dataset01

Figure 14 - Create Data Set

- Step 03 – Create Table

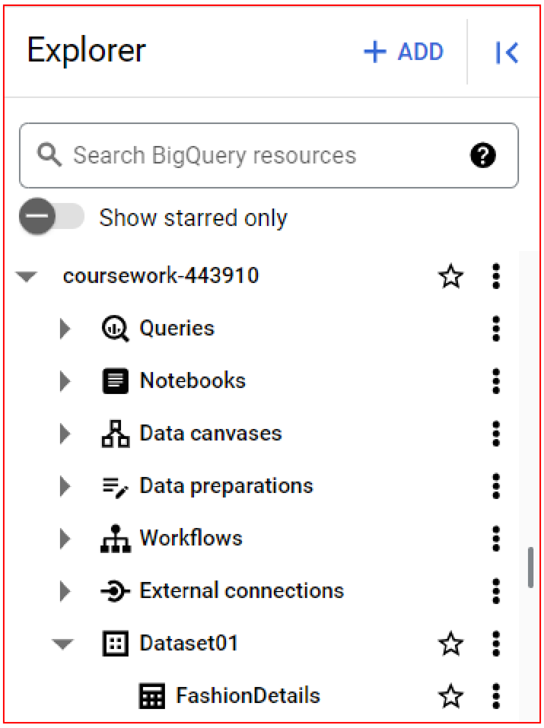
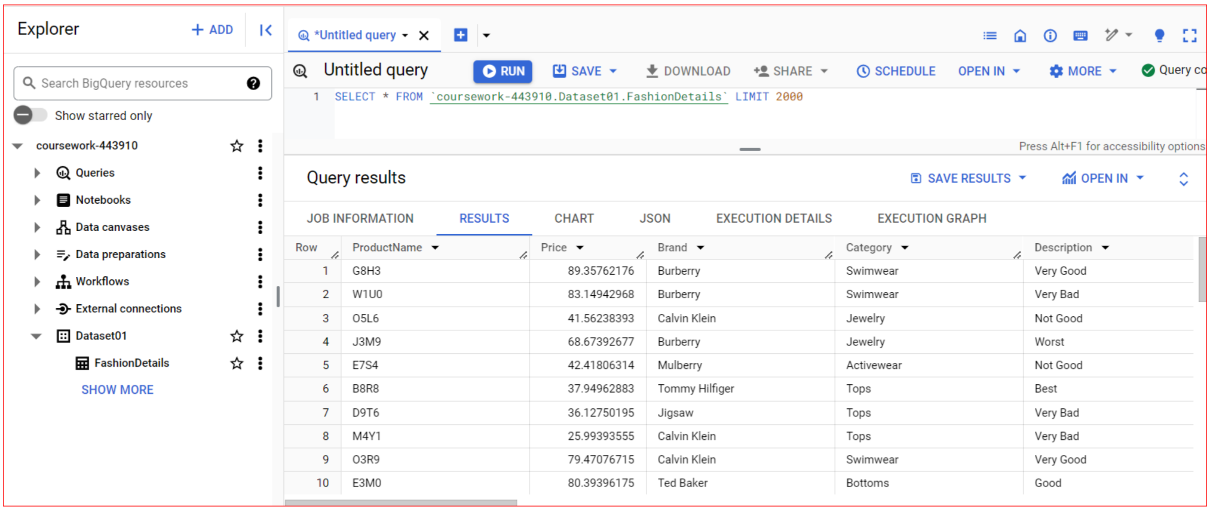


Figure 15 - Create Table

- Step 04 – Display All the Data

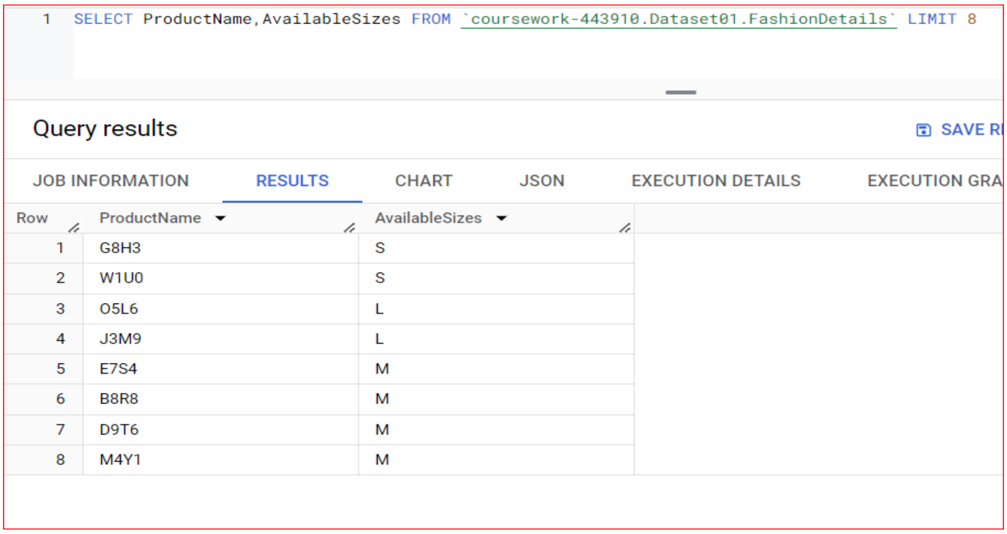


```
SELECT * FROM `coursework-443910.Dataset01.FashionDetails` LIMIT 2000
```

Row	ProductName	Price	Brand	Category	Description
1	G8H3	89.35762176	Burberry	Swimwear	Very Good
2	W1U0	83.14942968	Burberry	Swimwear	Very Bad
3	O5L6	41.56238393	Calvin Klein	Jewelry	Not Good
4	J3M9	68.67392677	Burberry	Jewelry	Worst
5	E7S4	42.41806314	Mulberry	Activewear	Not Good
6	B8R8	37.94962883	Tommy Hilfiger	Tops	Best
7	D9T6	36.12750195	Jigsaw	Tops	Very Bad
8	M4Y1	25.99393555	Calvin Klein	Tops	Very Bad
9	O3R9	79.47076715	Calvin Klein	Swimwear	Very Good
10	E3M0	80.39396175	Ted Baker	Bottoms	Good

Figure 16 - Display All the Data

- Step 05 – Display Columns Dataset



```
SELECT ProductName,AvailableSizes FROM `coursework-443910.Dataset01.FashionDetails` LIMIT 8
```

Row	ProductName	AvailableSizes
1	G8H3	S
2	W1U0	S
3	O5L6	L
4	J3M9	L
5	E7S4	M
6	B8R8	M
7	D9T6	M
8	M4Y1	M

Figure 17 - Display Column Dataset

3.3 Task 3: Connecting Tableau Desktop

- Step 01 – Go to Big Query

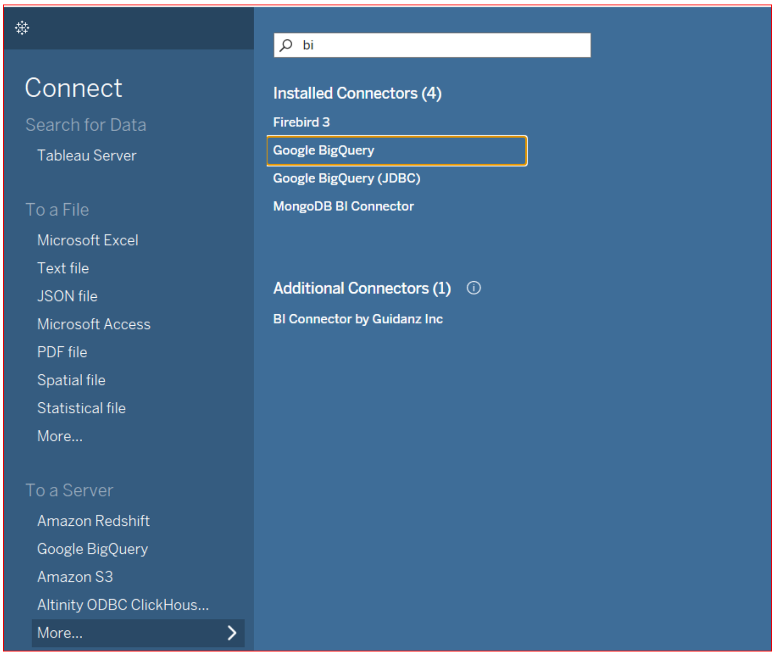
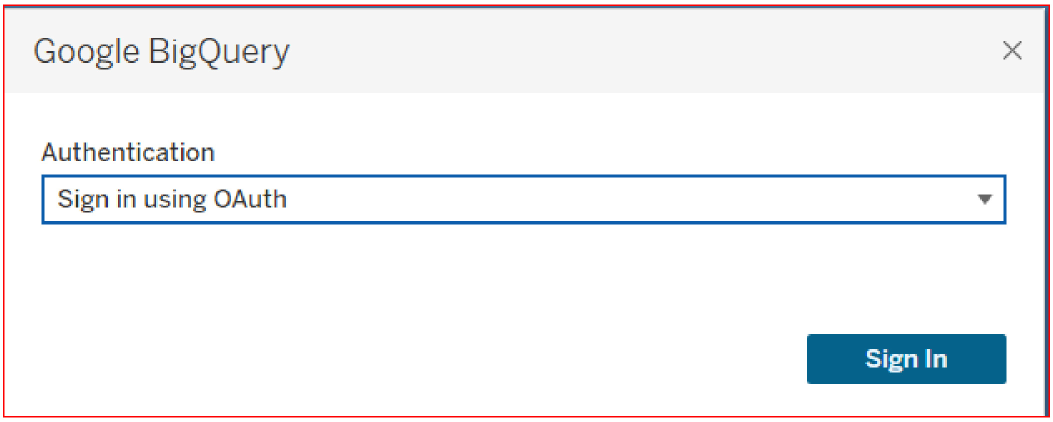


Figure 18 - Go to Big Query

- Step 02 – Connect to Big Query



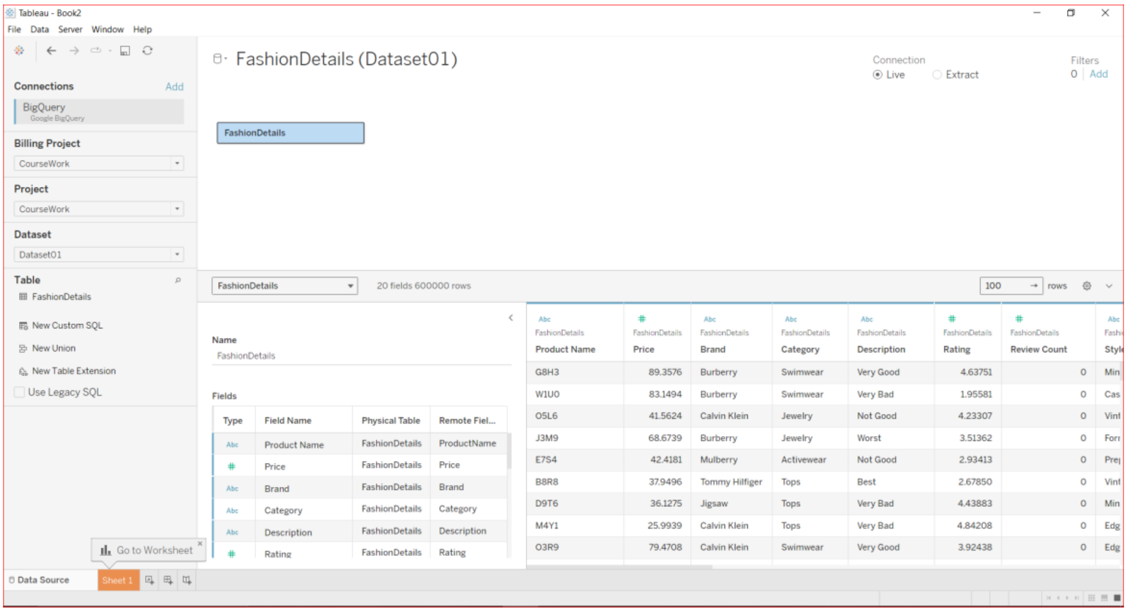
Google BigQuery

Authentication

Sign in using OAuth

Figure 19 - Login to Tableau

- Step 03 – Open Project



Billing Project: CourseWork

Project: CourseWork

Dataset: Dataset01

Table: FashionDetails

☐ Use Legacy SQL

FashionDetails (Dataset01)

Connection ● Live ○ Extract

Filters 0 | Add

FashionDetails

20 fields 600000 rows

100 rows

Name: FashionDetails

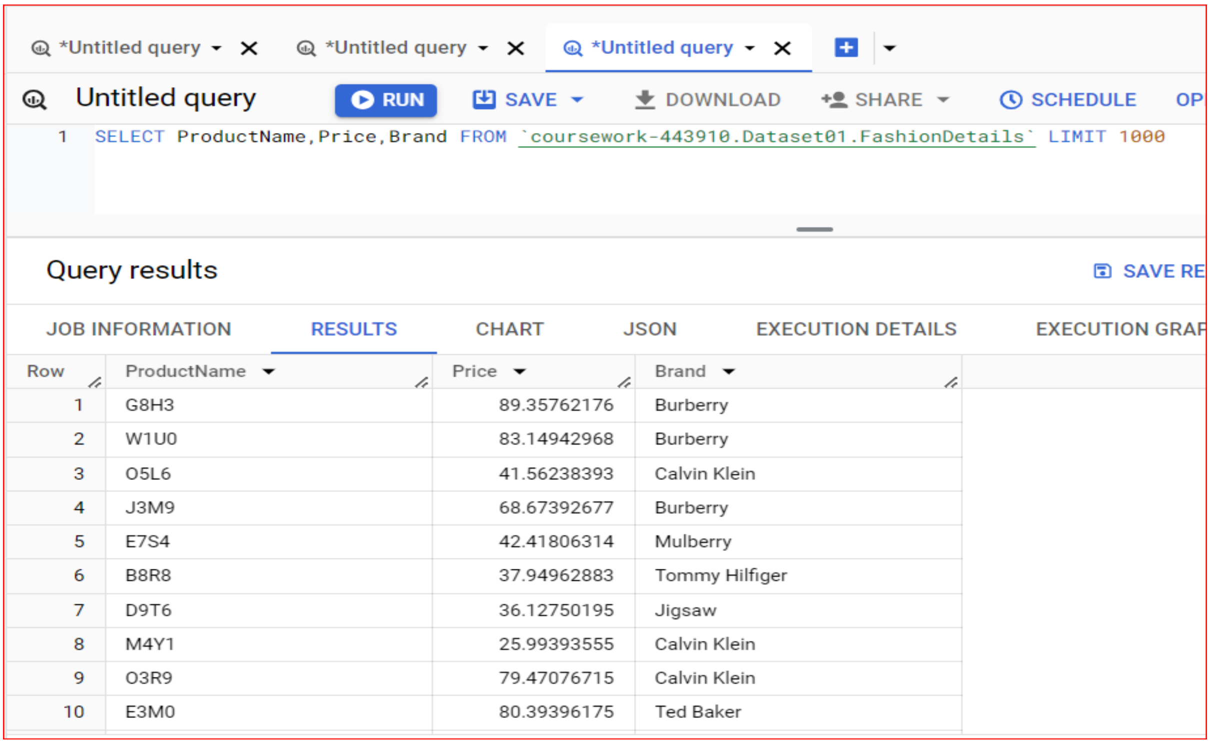
Fields

Type	Field Name	Physical Table	Remote Fiel...
Abc	Product Name	FashionDetails	ProductName
#	Price	FashionDetails	Price
Abc	Brand	FashionDetails	Brand
Abc	Category	FashionDetails	Category
Abc	Description	FashionDetails	Description
#	Rating	FashionDetails	Rating

Abc FashionDetails Product Name	# FashionDetails Price	Abc FashionDetails Brand	Abc FashionDetails Category	Abc FashionDetails Description	# FashionDetails Rating	# FashionDetails Review Count	Abc Fashi Styl
G8H3	89.3576	Burberry	Swimwear	Very Good	4.63751	0	Min
W1U0	83.1494	Burberry	Swimwear	Very Bad	1.95581	0	Cas
O5L6	41.5624	Calvin Klein	Jewelry	Not Good	4.23307	0	Vint
J3M9	68.6739	Burberry	Jewelry	Worst	3.51362	0	Forr
E7S4	42.4181	Mulberry	Activewear	Not Good	2.93413	0	Pre
B8R8	37.9496	Tommy Hilfiger	Tops	Best	2.67850	0	Vint
D9T6	36.1275	Jigsaw	Tops	Very Bad	4.43883	0	Min
M4Y1	25.9939	Calvin Klein	Tops	Very Bad	4.84208	0	Edg
O3R9	79.4708	Calvin Klein	Swimwear	Very Good	3.92438	0	Edg

Figure 20 - Open Project and View Table

- Step 04 – Display Details about product name their prices and brands

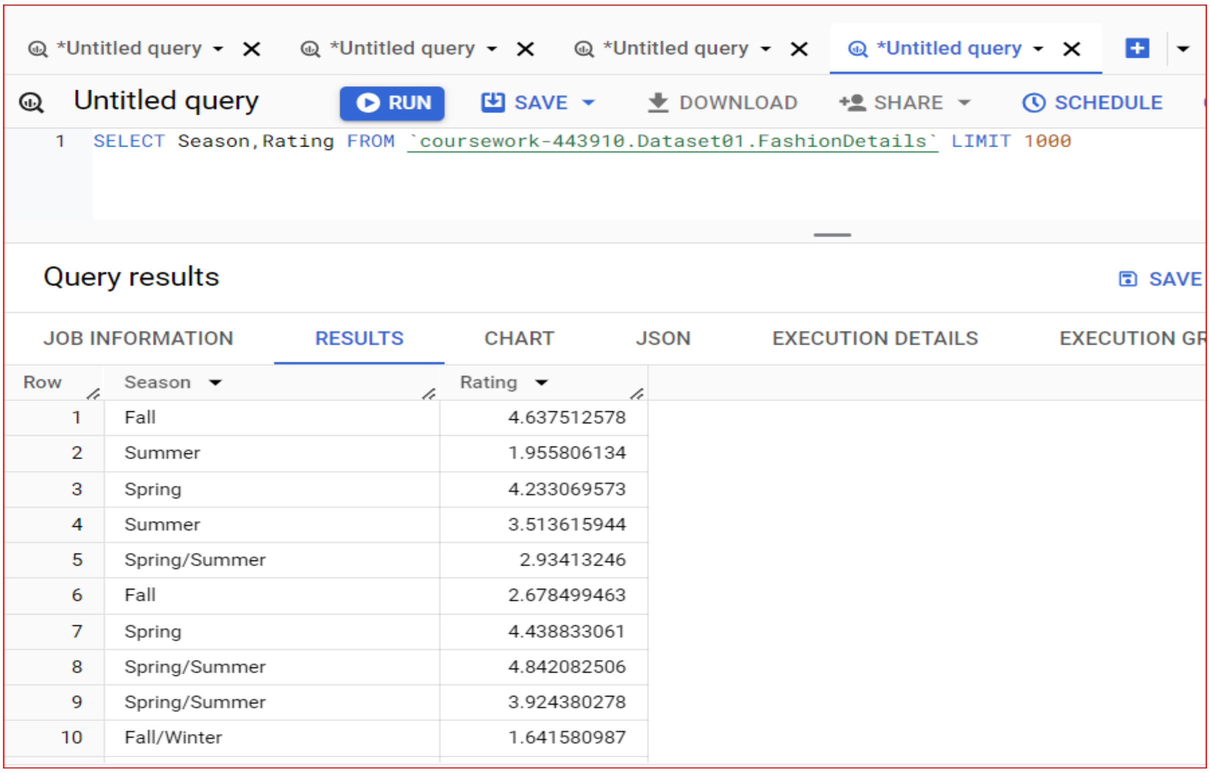


```
SELECT ProductName,Price,Brand FROM `coursework-443910.Dataset01.FashionDetails` LIMIT 1000
```

Row	ProductName	Price	Brand
1	G8H3	89.35762176	Burberry
2	W1U0	83.14942968	Burberry
3	O5L6	41.56238393	Calvin Klein
4	J3M9	68.67392677	Burberry
5	E7S4	42.41806314	Mulberry
6	B8R8	37.94962883	Tommy Hilfiger
7	D9T6	36.12750195	Jigsaw
8	M4Y1	25.99393555	Calvin Klein
9	O3R9	79.47076715	Calvin Klein
10	E3M0	80.39396175	Ted Baker

Figure 21 - Display Details about Name, Price, Brands

- Step 04 – Display Data Rating of Seasons



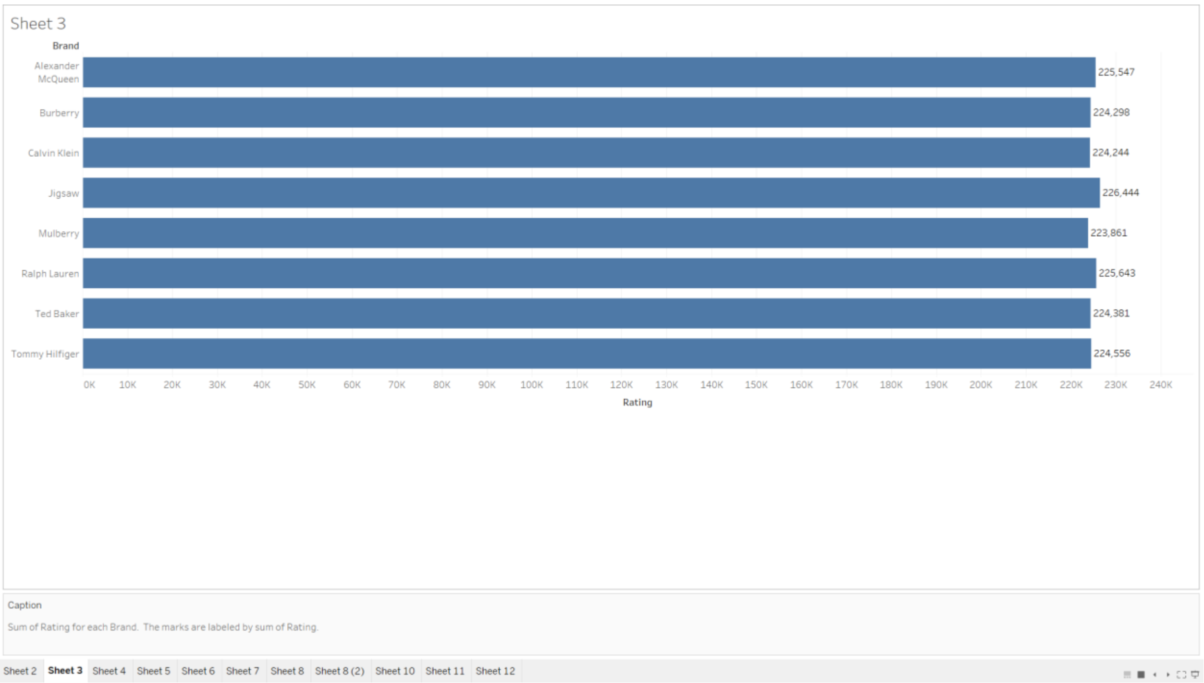
```
SELECT Season,Rating FROM `coursework-443910.Dataset01.FashionDetails` LIMIT 1000
```

Row	Season	Rating
1	Fall	4.637512578
2	Summer	1.955806134
3	Spring	4.233069573
4	Summer	3.513615944
5	Spring/Summer	2.93413246
6	Fall	2.678499463
7	Spring	4.438833061
8	Spring/Summer	4.842082506
9	Spring/Summer	3.924380278
10	Fall/Winter	1.641580987

Figure 22 - Display Rating of Seasons

3.4 Task 4: Creating Data Visualizations in Tableau Desktop

- Sum of Rating for Each Brands.



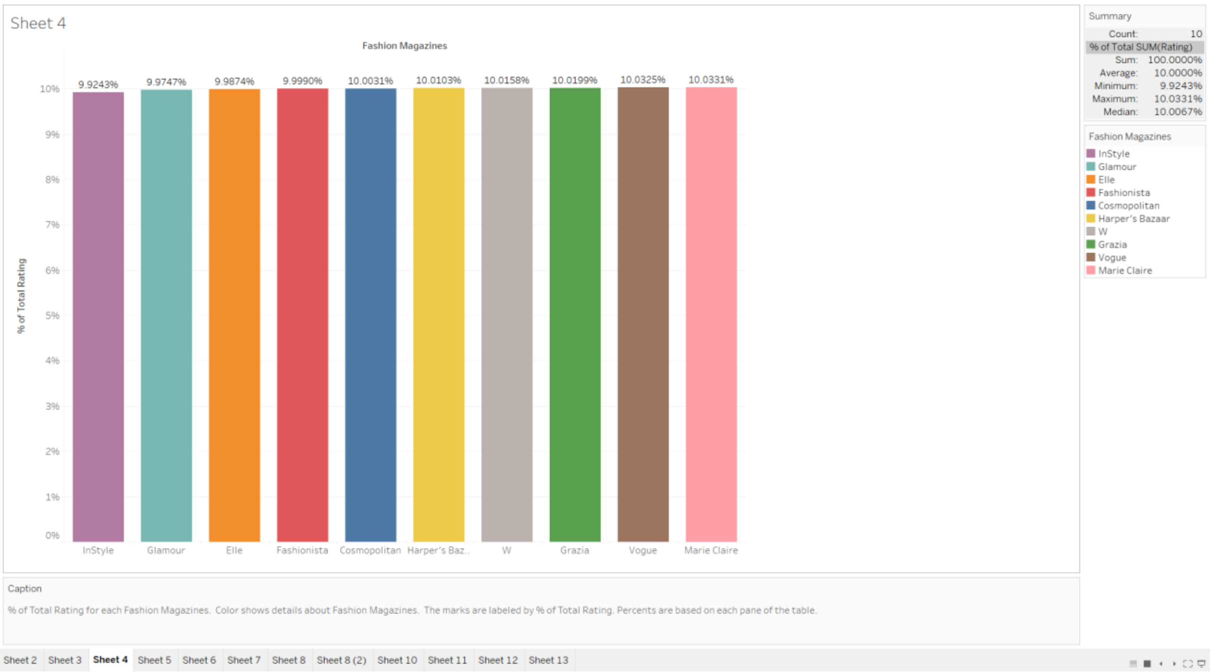
Caption

Sum of Rating for each Brand. The marks are labeled by sum of Rating.

Figure 23 - Sum of Ration for Each Brands

- Percentage of Total Rating for each Fashion Magazines.

Color shows details about Fashion Magazines. The marks are labeled by & of Total Rating. Precents are based on each pane of the table



Summary

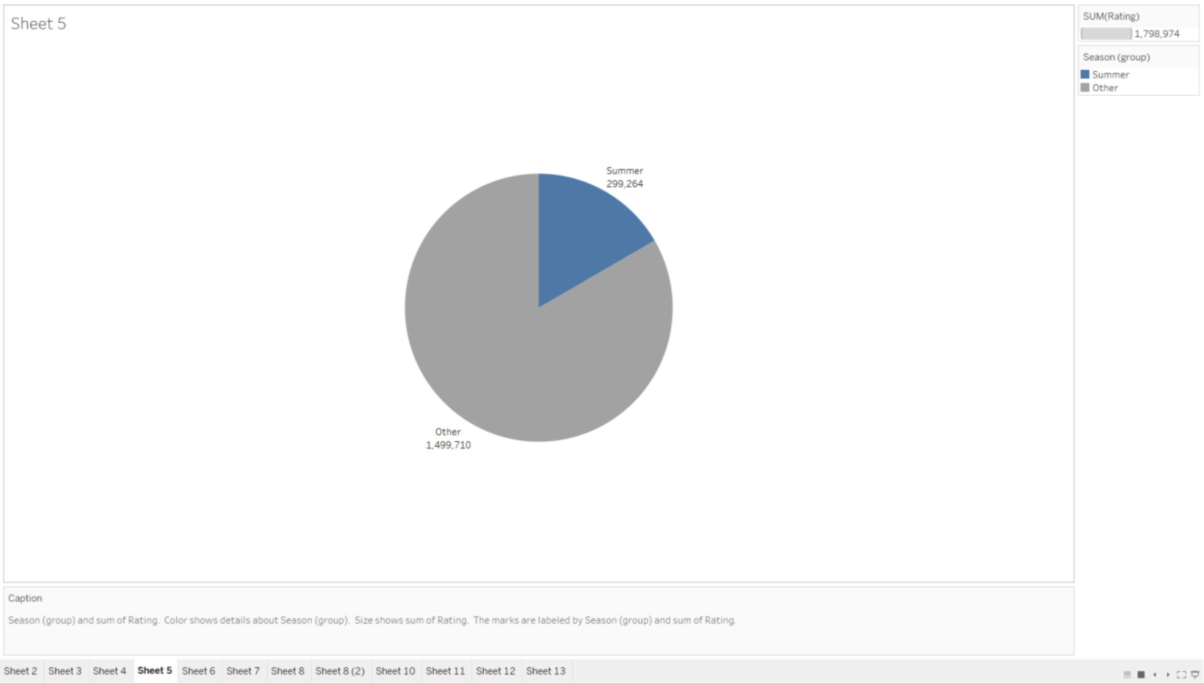
Count:	10
% of Total SUM(Rating)	
Sum:	100.0000%
Average:	10.0000%
Minimum:	9.9243%
Maximum:	10.0331%
Median:	10.0067%

Caption

% of Total Rating for each Fashion Magazines. Color shows details about Fashion Magazines. The marks are labeled by % of Total Rating. Percents are based on each pane of the table.

Figure 24 - Percentage of Total Rating for each Fashion Magazines

- Season of Rating



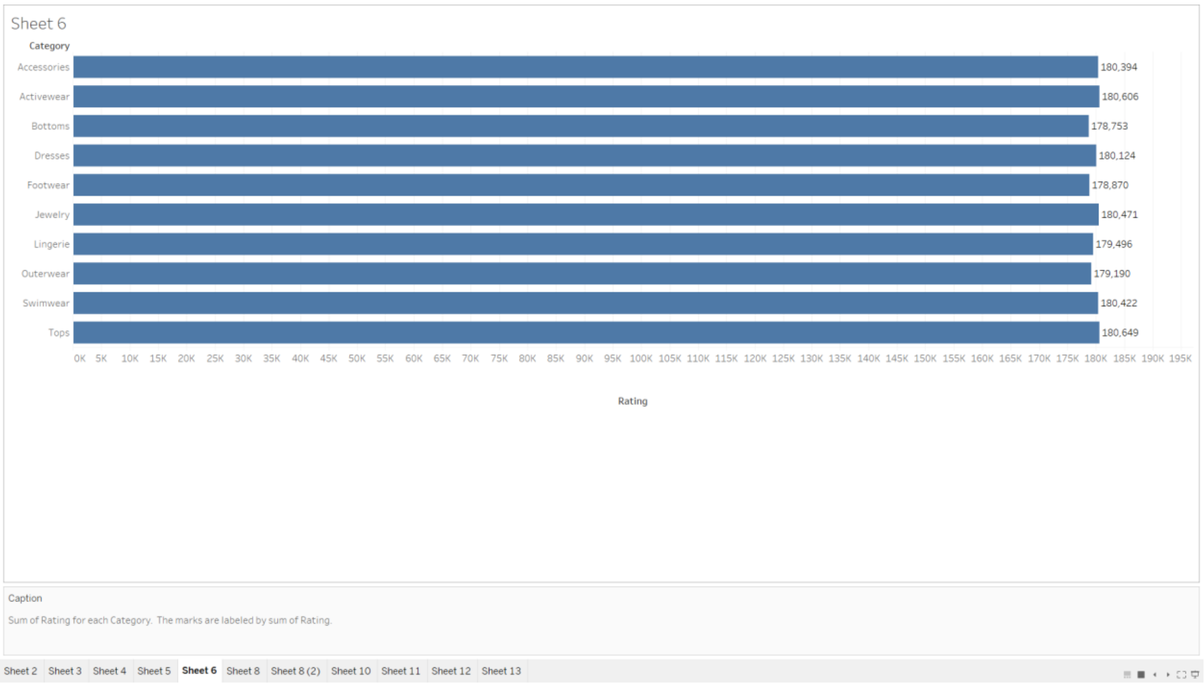
Caption

Season (group) and sum of Rating. Color shows details about Season (group). Size shows sum of Rating. The marks are labeled by Season (group) and sum of Rating.

Figure 25 - Season of Rating

- Sum of Rating for each Category

Sum of Rating for each Categories are Accessories, Bottoms, Activewear, Dresses, Jewelry, Lingerie, Outerwear, Swimwear, Tops



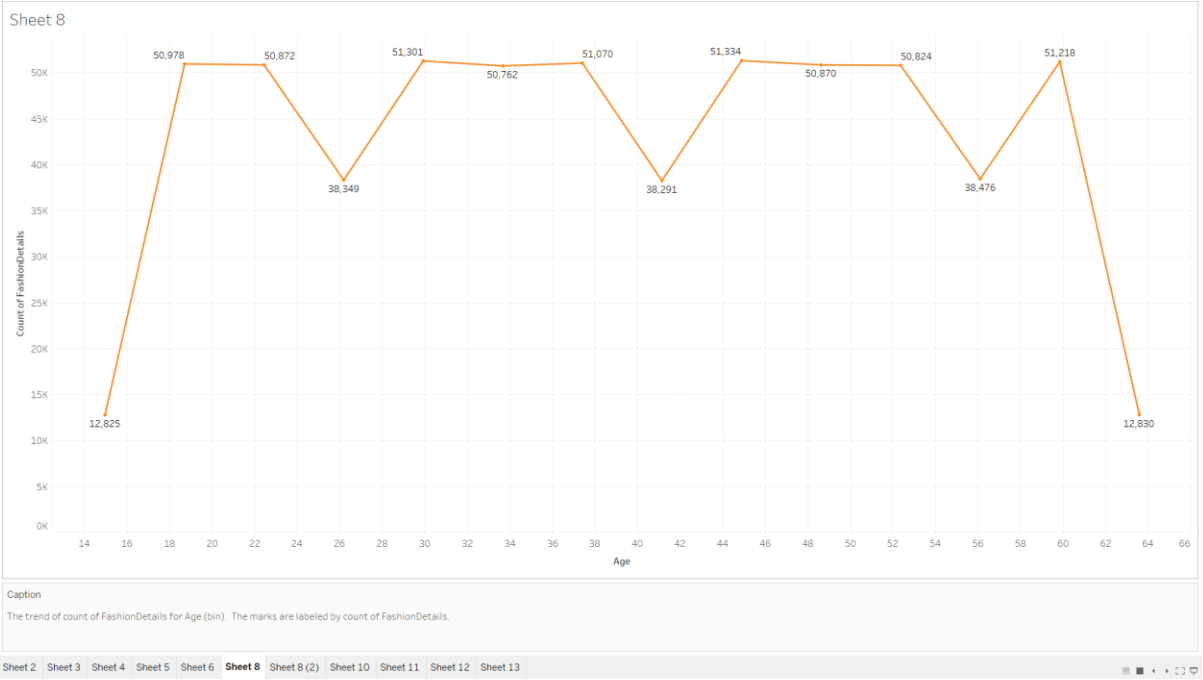
Caption

Sum of Rating for each Category. The marks are labeled by sum of Rating.

Figure 26 - Sum of Rating for each Category

- The trend of count of Fashion Details for Age

The trend of count of Fashion Details for Age. The marks are labeled by count of Fashion Details.



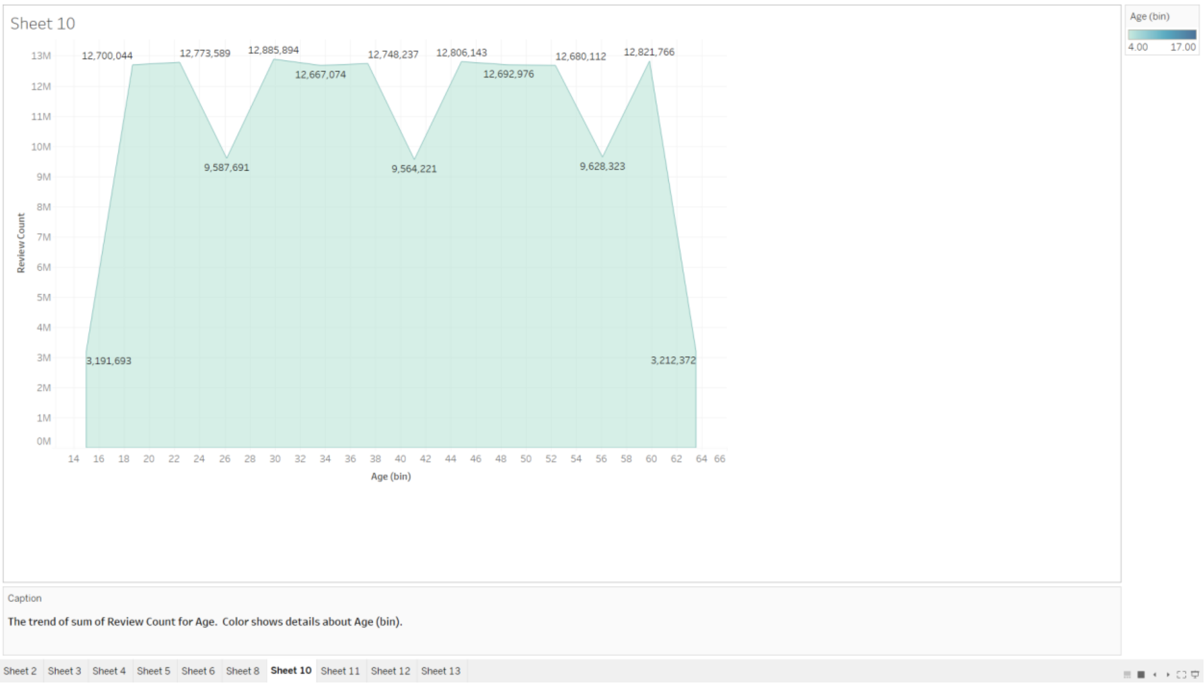
Caption

The trend of count of FashionDetails for Age (bin). The marks are labeled by count of FashionDetails.

Figure 27 - The trend of count of Fashion Details for Age

- The trend of sum of Review Count for Age

The trend of sum of Review Count for Age. Color shows details about Age.



Caption

The trend of sum of Review Count for Age. Color shows details about Age (bin).

Figure 28 - The trend of sum of Review Count for Age

4.0 DISCUSSION

First, as mentioned in Coursework - C, I completed the first task of the lecturer. That is, I logged in to the oracle account and created the Autonomous Datawarehouse and connected to the Oracle SQL developer, but the Autonomous Datawarehouse could not connect to the tableau. So, I used Amazon Redshift to connect the database to the tableau, but the tableau could not do it. So, I tried to upload the database to Google Big Query. I used Google Colab to clean the data and used a Python script for that. For that, I used pandas and since the data size needed to be less than 100mb, I cleaned the data and removed the last 400000 rows. Then the data size could be reduced to less than 100mb. The (csv) Excel file was uploaded to the Google Big Query dataset. After that, I typed the required queries and created the Data warehouse. In the data cleaning part, I selected a dataset from Kaggle and then uploaded that dataset to Google Colab and then used Python code to clean it. Then, after connecting Google Big Query to Tableau, the visualization part was done through it. For that, pie chart, bar chart, Area chart was used.

5.0 CONCLUTION

The Fashion Sales Dataset UK-US project provides a comprehensive view of the fashion industry’s trends and sales. The meticulously curated dataset serves as a reliable resource for analyzing sales trends, forecasting demand, and optimizing business strategies in the UK and US markets By incorporating a variety of attributes such as product descriptions, customer interactions, and seasonal variations, this dataset enables researchers, analysts, and industry professionals to gain actionable insights into consumer behavior and market dynamics.

This dataset is a powerful tool for identifying market trends, understanding the impact of seasonal preferences, and uncovering the role of external factors such as fashion influencers and social media. It is designed to help fashion industry decision-makers make data-driven decisions that can improve sales strategies, product development, and customer engagement.

Overall, the Fashion Sales Dataset bridges the gap between sales data and strategic decision-making, providing a solid foundation for informed planning and innovation in the dynamic fashion industry.

6.0 REFERENCES

1. **Connect Tableau**
    - https://help.tableau.com/
    - Use this to connect Tableau to Google Big Query to analysis and Visualization dataset

2. **Query & Create table in Google Big Query**
    - To find create dataset, create table, csv file upload and view dataset
    - https://cloud.google.com

3. **Kaggle**
    - To Download dataset
    - https://www.kaggle.com/