****

**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

[1.0 INTRODUCTION 5](#_Toc184553939)

[2.0 PREREQUISITES 6](#_Toc184553940)

[2.1 Required Software and Tools 6](#_Toc184553941)

[3.0 COURSEWORK TASKS 7](#_Toc184553942)

[3.1 Task 1: Sourcing and Preparing Sample Data 7](#_Toc184553943)

[3.2 Task 2: Amazon Redshift Account Create for Cloud Data Warehousing 11](#_Toc184553944)

[3.3 Task 3: Connecting Tableau Desktop 13](#_Toc184553945)

[3.4 Task 4: Creating Data Visualizations in Tableau Desktop 16](#_Toc184553946)

[4.0 DISCUSSION 19](#_Toc184553947)

[5.0 CONCLUTION 19](#_Toc184553948)

[6.0 REFERENCES 20](#_Toc184553949)

Table of Figures

[Figure 01- Uncleared data set 7](#_Toc184553911)

[Figure 02 - Import Pandas version 8](#_Toc184553912)

[Figure 03 - Display Pandas Version 8](#_Toc184553913)

[Figure 04 - Read Unclear Dataset Data 8](#_Toc184553914)

[Figure 05 - Drops Rows with Missing (Non) Values 8](#_Toc184553915)

[Figure 06 - Remove Duplicate Values 8](#_Toc184553916)

[Figure 07 - Display Row Count Before Clean 9](#_Toc184553917)

[Figure 08 - Clean Data Column 9](#_Toc184553918)

[Figure 09 - Remove Rows for Reduce the File Size 9](#_Toc184553919)

[Figure 10 - Display Last Modified Data Set 9](#_Toc184553920)

[Figure 11 - Download Clean Dataset 10](#_Toc184553921)

[Figure 12 - Clean Dataset Size 10](#_Toc184553922)

[Figure 13 - Create New Project 11](#_Toc184553923)

[Figure 14 - Create Data Set 11](#_Toc184553924)

[Figure 15 - Create Table 12](#_Toc184553925)

[Figure 16 - Display All the Data 12](#_Toc184553926)

[Figure 17 - Display Column Dataset 13](#_Toc184553927)

[Figure 18 - Go to Big Query 13](#_Toc184553928)

[Figure 19 - Login to Tableau 14](#_Toc184553929)

[Figure 20 - Open Project and View Table 14](#_Toc184553930)

[Figure 21 - Display Details about Name, Price, Brands 15](#_Toc184553931)

[Figure 22 - Display Rating of Seasons 15](#_Toc184553932)

[Figure 23 - Sum of Ration for Each Brands 16](#_Toc184553933)

[Figure 24 - Percentage of Total Rating for each Fashion Magazines 16](#_Toc184553934)

[Figure 25 - Season of Rating 17](#_Toc184553935)

[Figure 26 - Sum of Rating for each Category 17](#_Toc184553936)

[Figure 27 - The trend of count of Fashion Details for Age 18](#_Toc184553937)

[Figure 28 - The trend of sum of Review Count for Age 18](#_Toc184553938)

# 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.

# PREREQUISITES

## 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

# COURSEWORK TASKS

## Task 1: Sourcing and Preparing Sample Data

1. Select Dataset

Here, a dataset of over 100MB was selected.

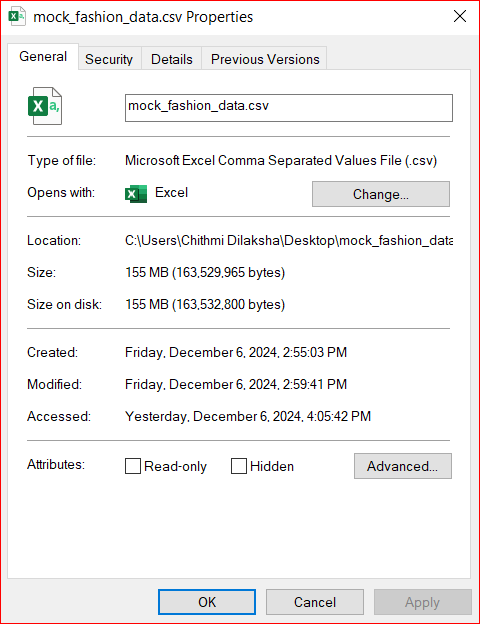


Figure 01- Uncleared data set

1. 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

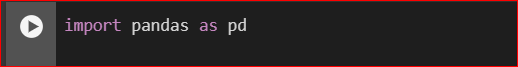


Figure 02 - Import Pandas version

* Step 02

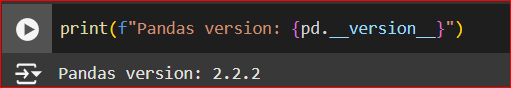


Figure 03 - Display Pandas Version

* Step 03

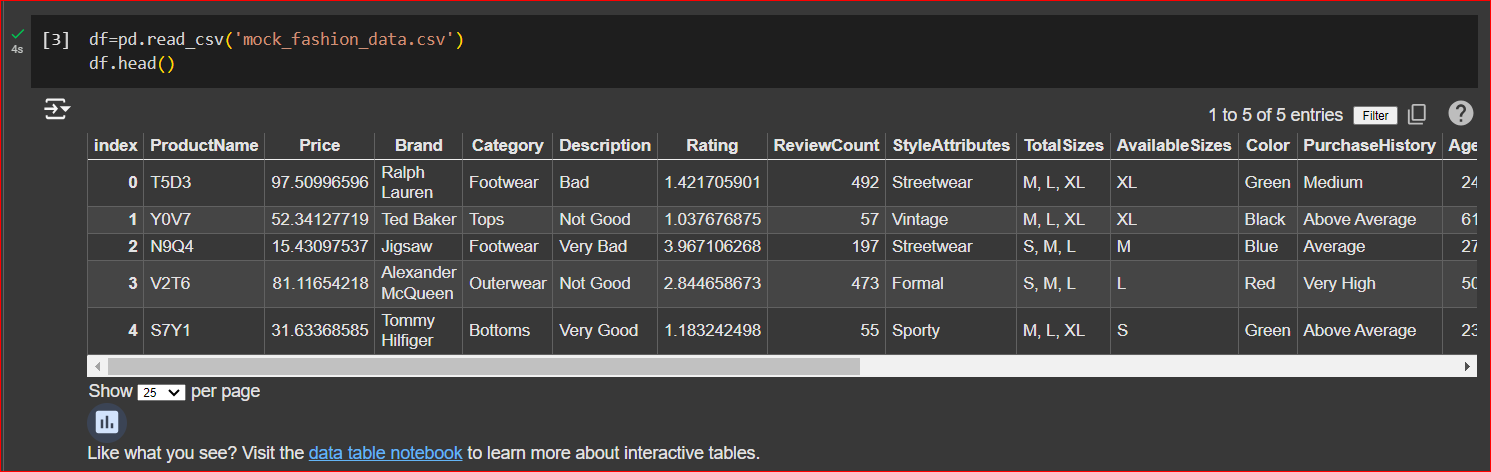


Figure 04 - Read Unclear Dataset Data

* Step 04



Figure 05 - Drops Rows with Missing (Non) Values

* Step 05



Figure 06 - Remove Duplicate Values

* Step 06

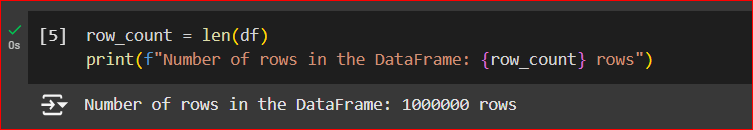


Figure 07 - Display Row Count Before Clean

* Step 07

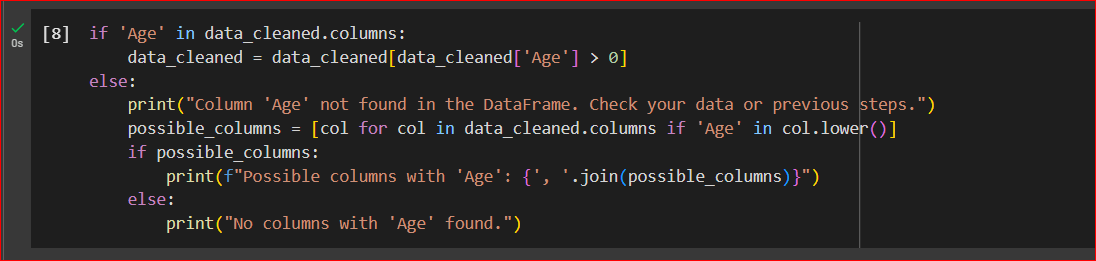


Figure 08 - Clean Data Column

* Step 08

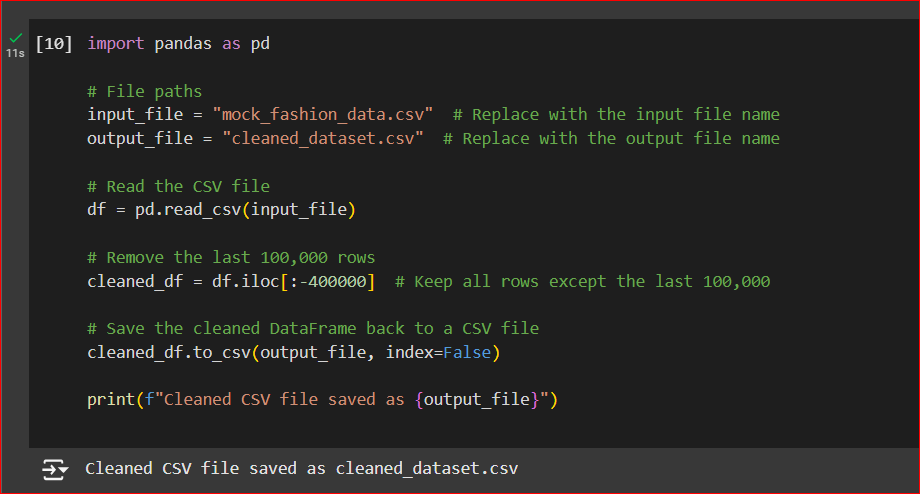


Figure 09 - Remove Rows for Reduce the File Size

* Step 09

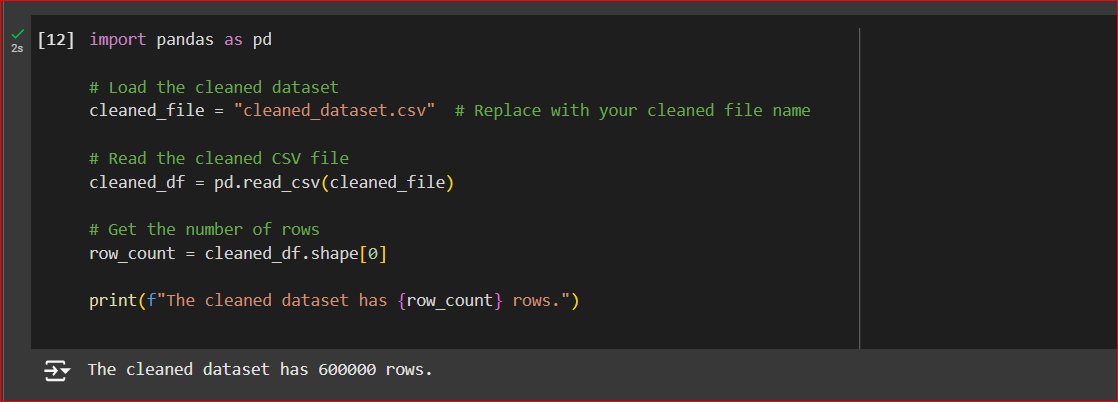


Figure 10 - Display Last Modified Data Set

* Step 10

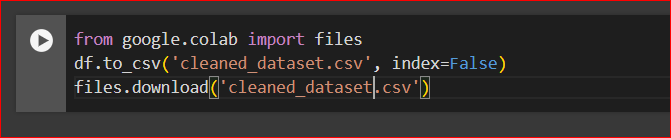


Figure 11 - Download Clean Dataset

1. After Cleaned Data Set

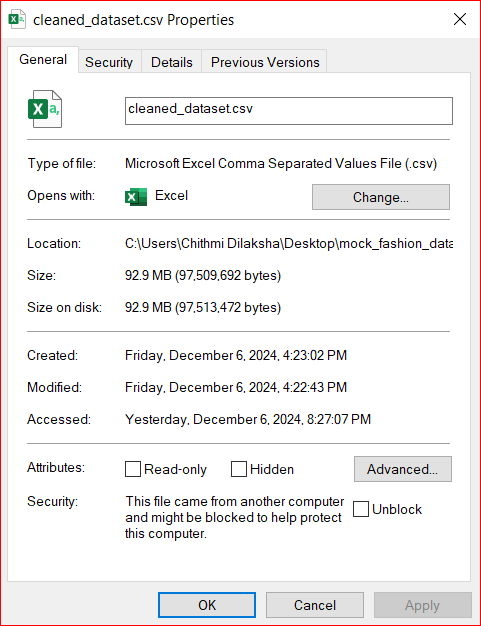


Figure 12 - Clean Dataset Size

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

* Step 01 – Create New Project

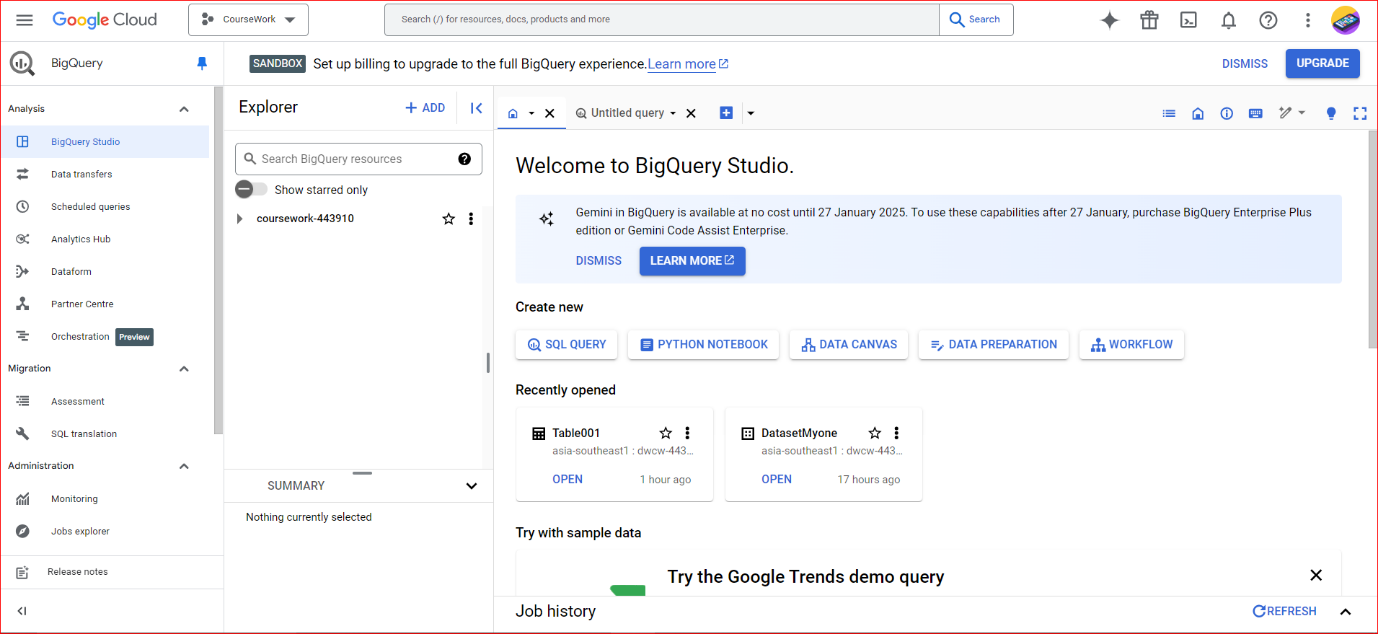


Figure 13 - Create New Project

* Step 02 – Create Data Set

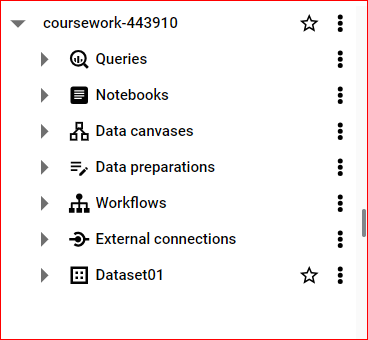


Figure 14 - Create Data Set

* Step 03 – Create Table

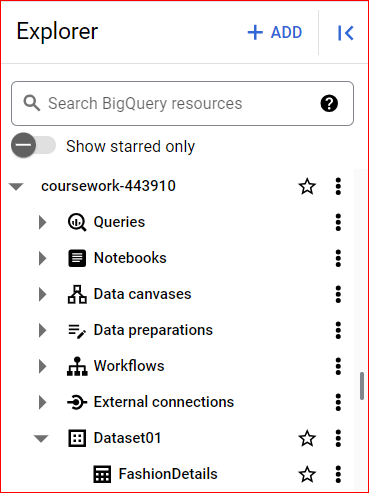


Figure 15 - Create Table

* Step 04 – Display All the Data

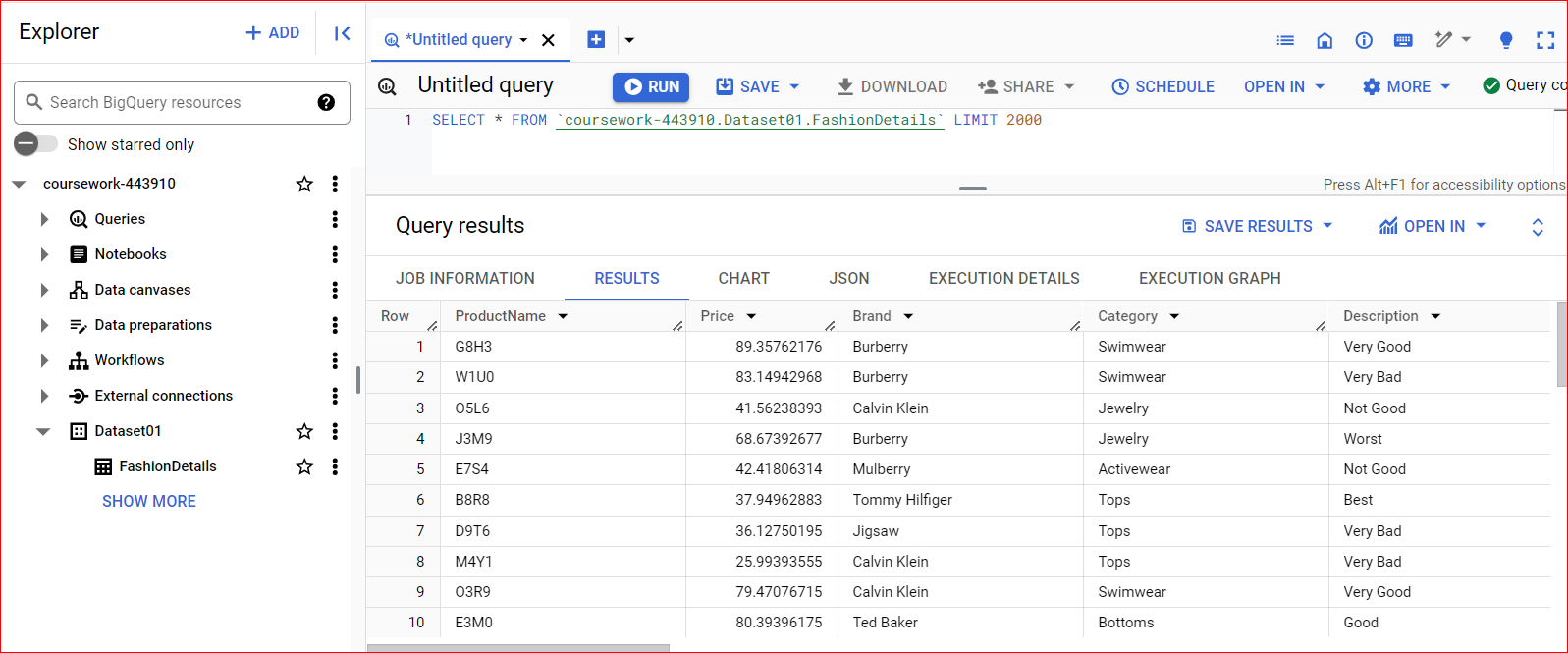


Figure 16 - Display All the Data

* Step 05 – Display Columns Dataset

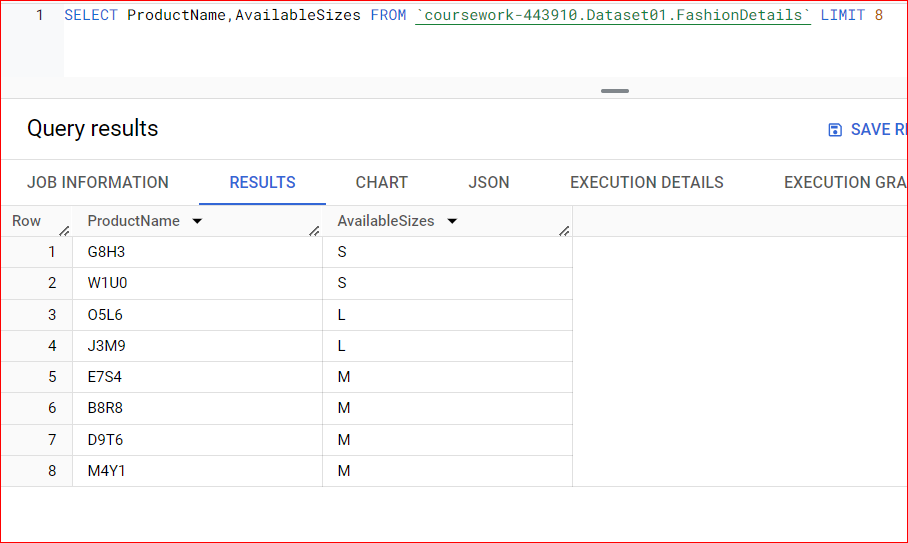


Figure 17 - Display Column Dataset

## Task 3: Connecting Tableau Desktop

* Step 01 – Go to Big Query

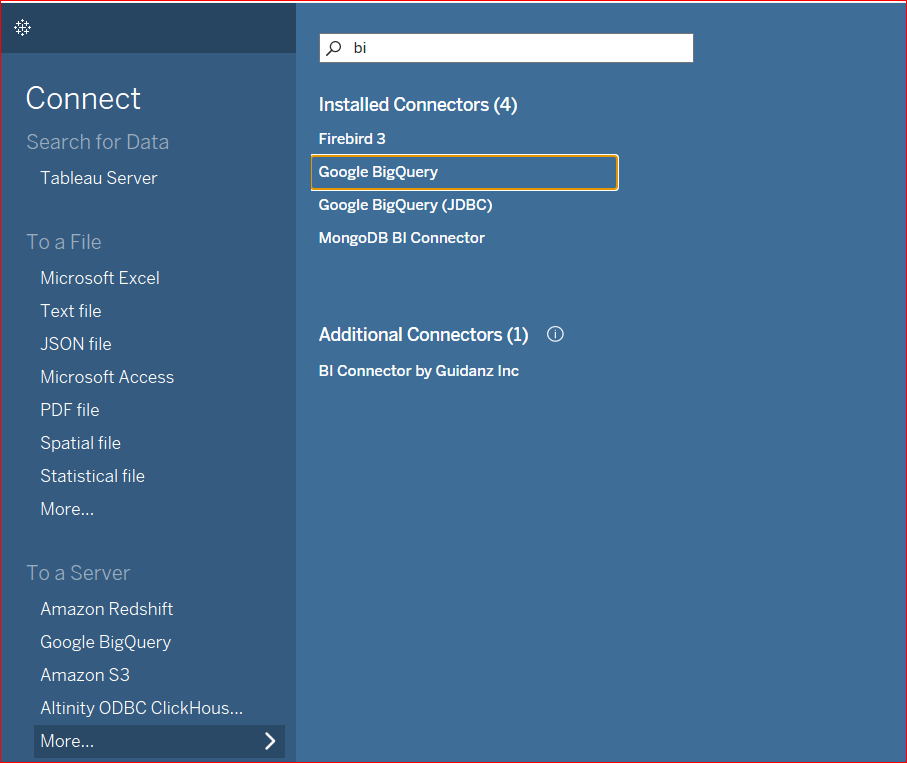


Figure 18 - Go to Big Query

* Step 02 – Connect to Big Query

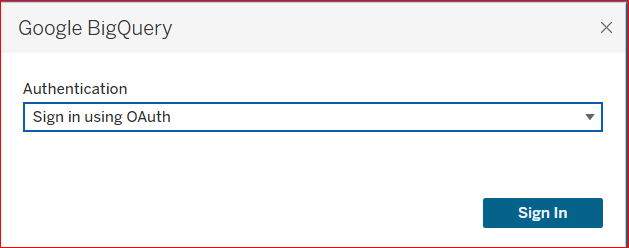


Figure 19 - Login to Tableau

* Step 03 – Open Project

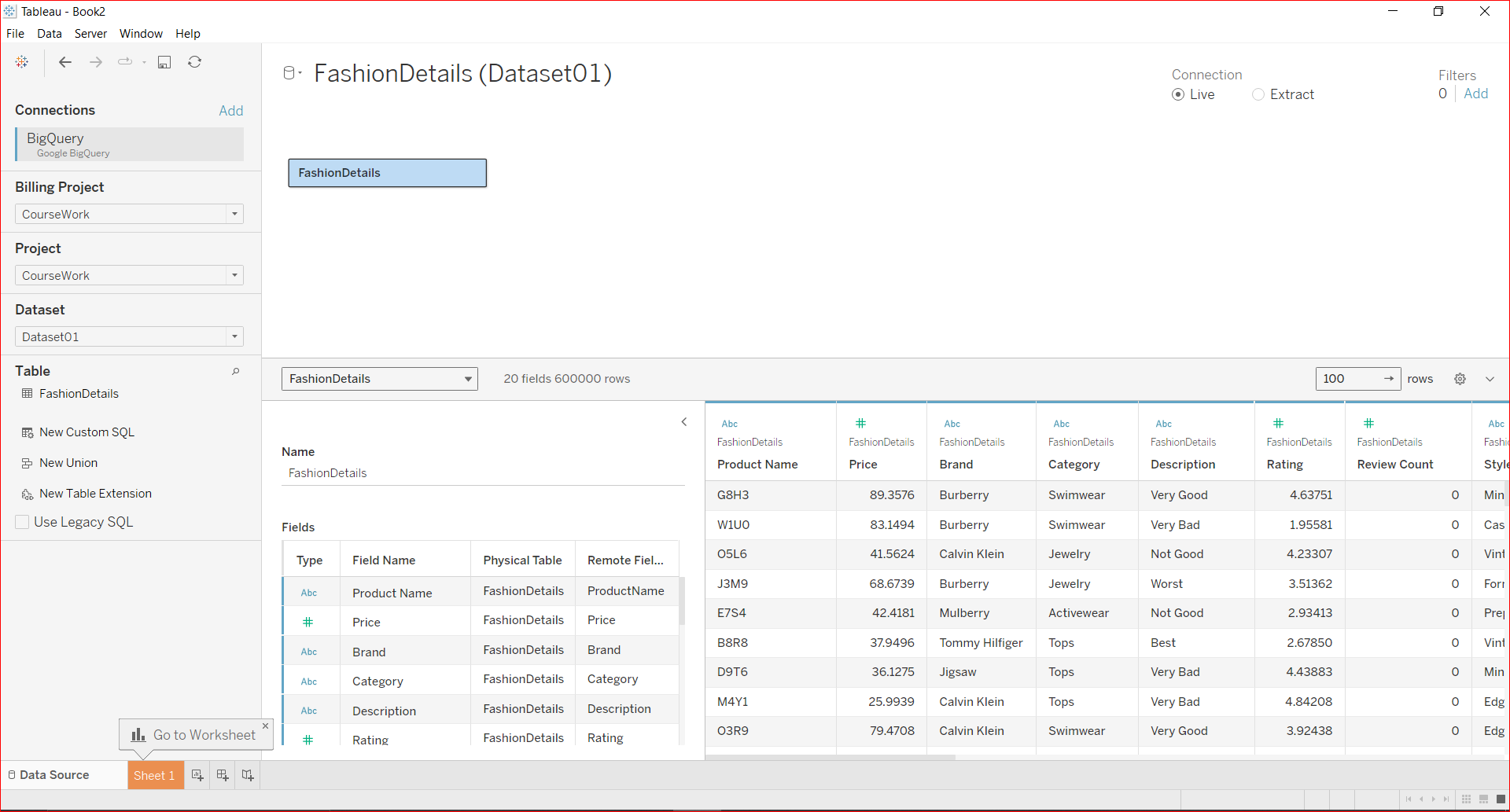


Figure 20 - Open Project and View Table

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

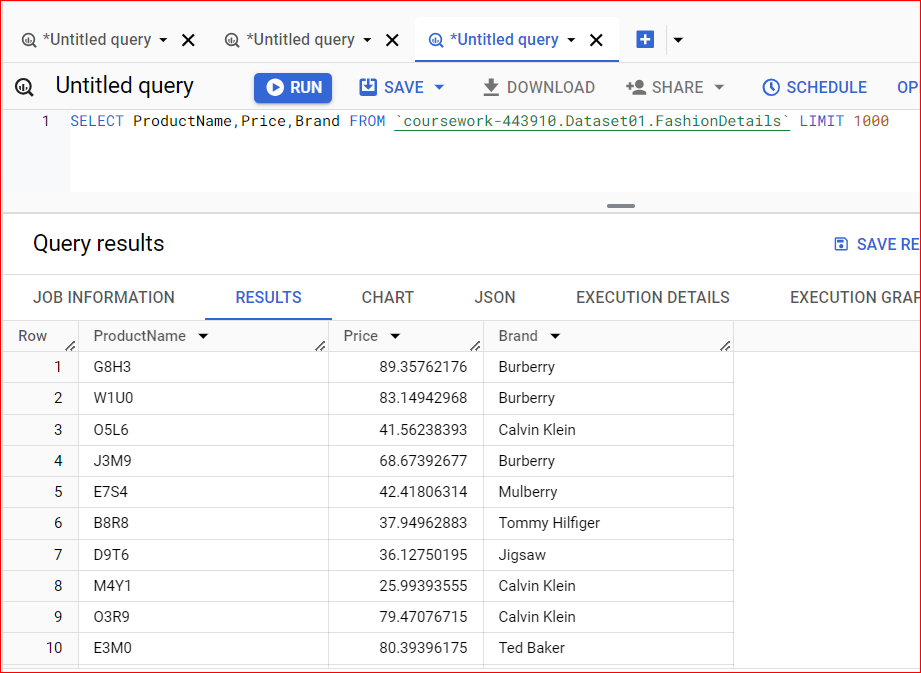


Figure 21 - Display Details about Name, Price, Brands

* Step 04 – Display Data Rating of Seasons

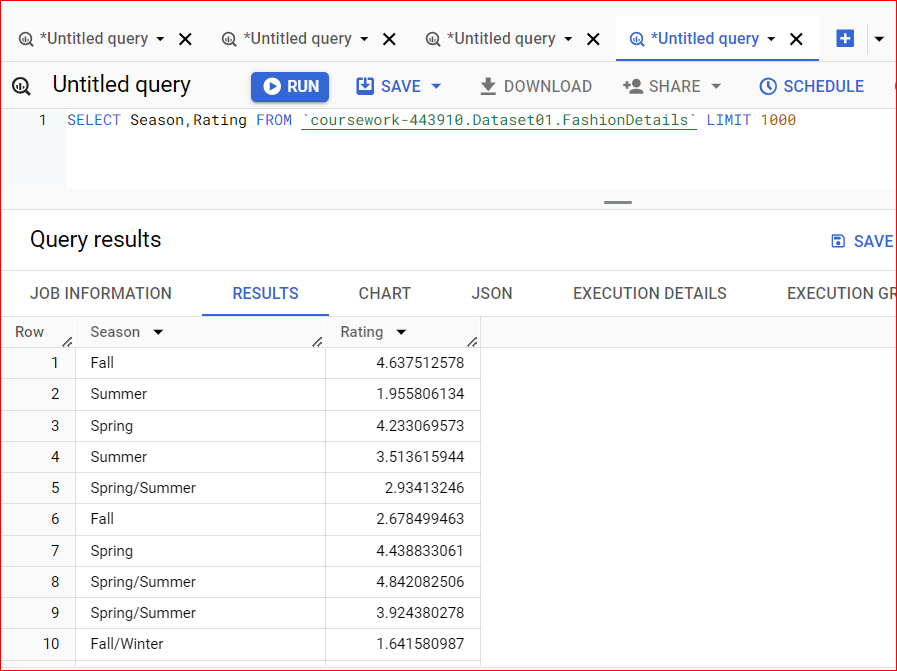


Figure 22 - Display Rating of Seasons

## Task 4: Creating Data Visualizations in Tableau Desktop

* Sum of Rating for Each Brands.

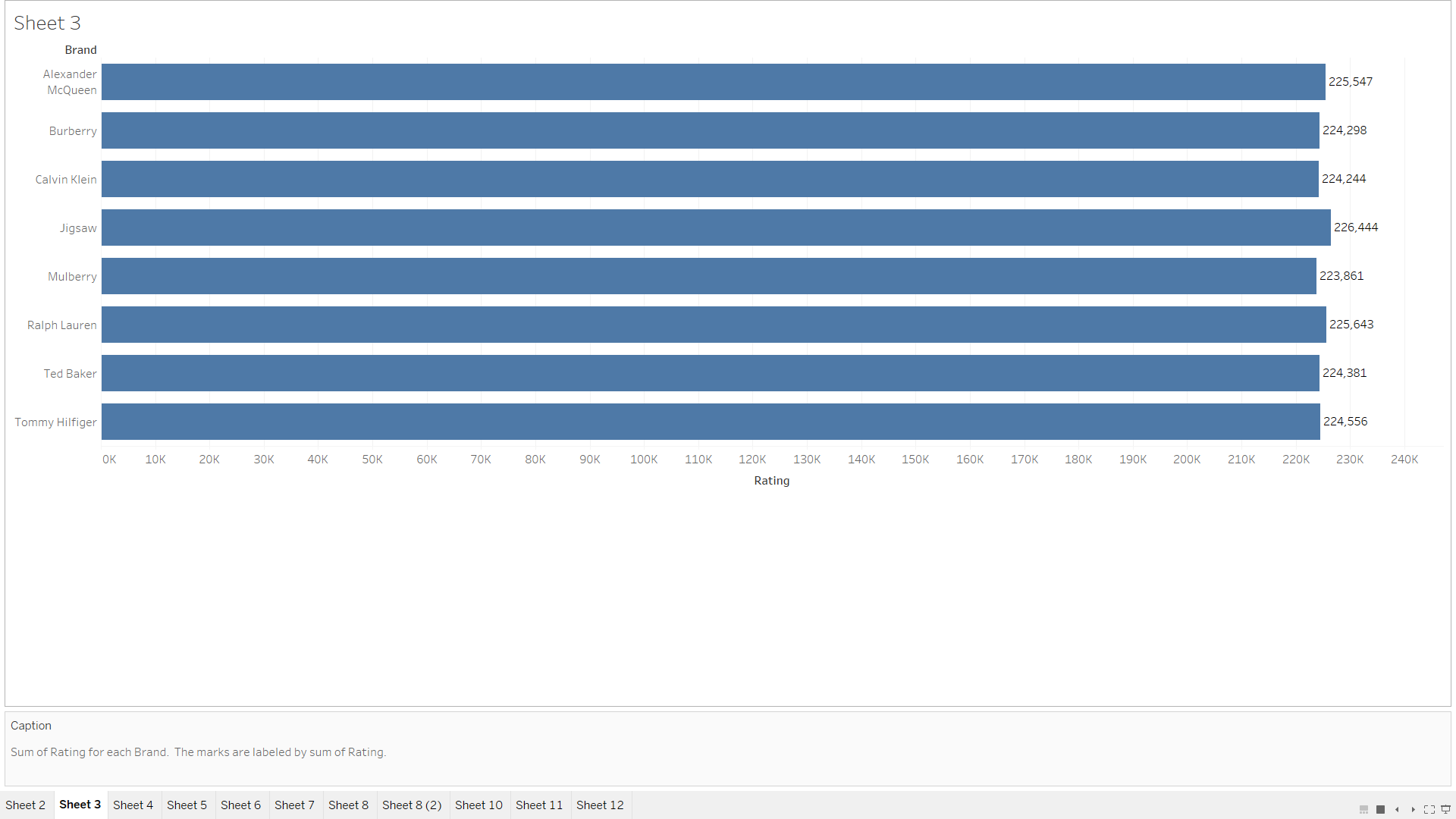


Figure 23 - Sum of Ration for Each Brands

* 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



Figure 24 - Percentage of Total Rating for each Fashion Magazines

* Season of Rating

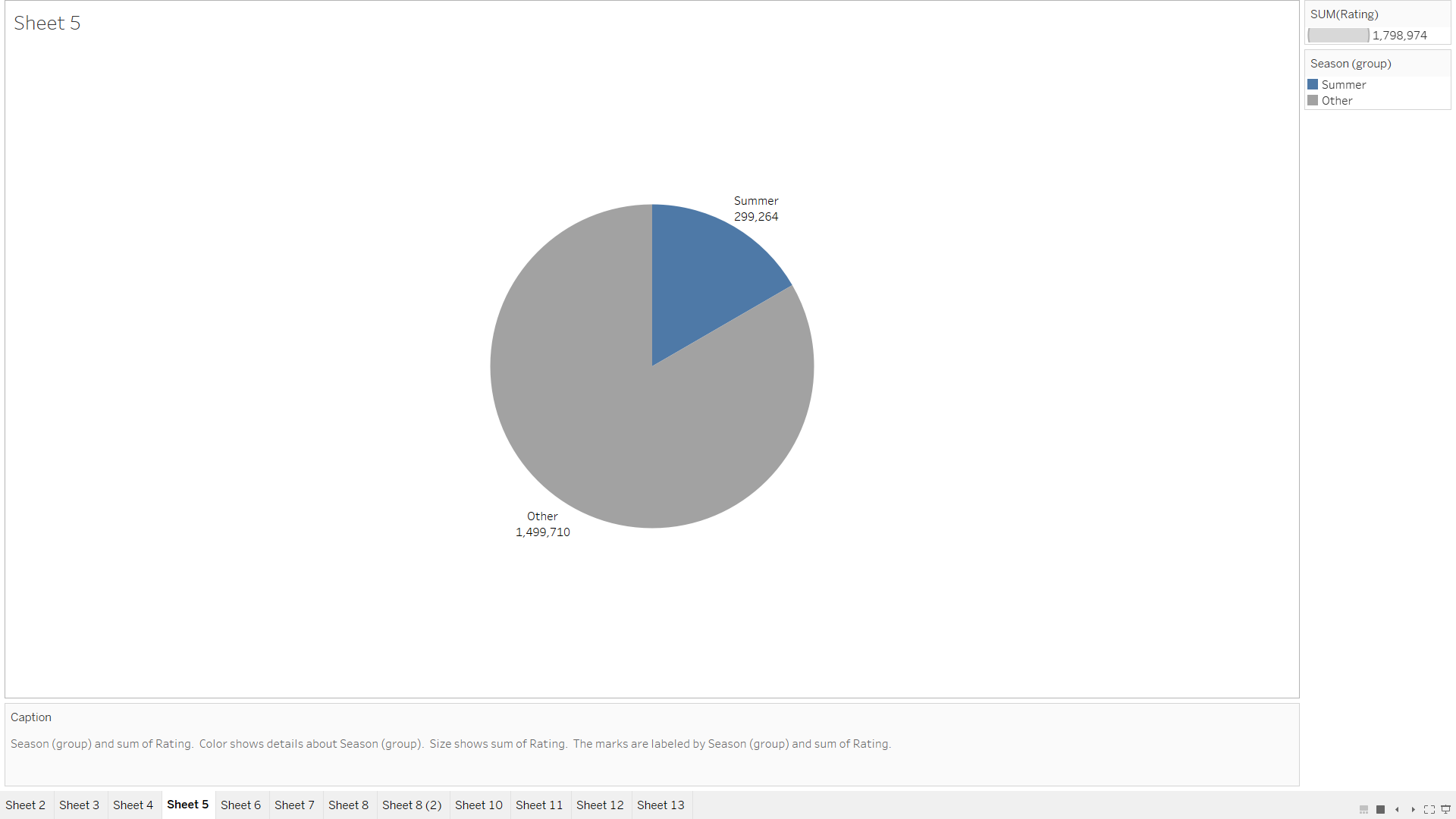


Figure 25 - Season of Rating

* Sum of Rating for each Category

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

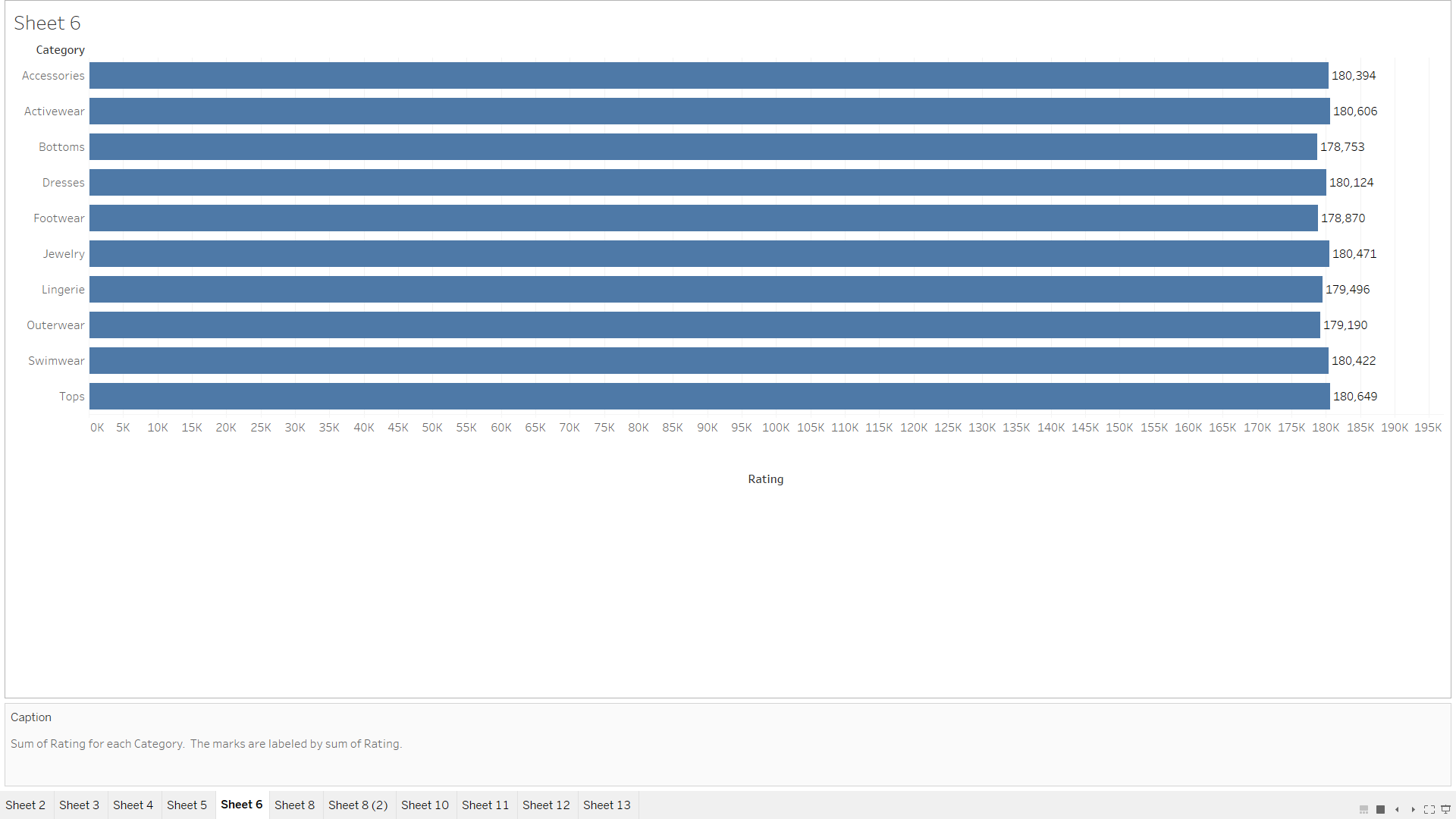


Figure 26 - Sum of Rating for each Category

* 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.

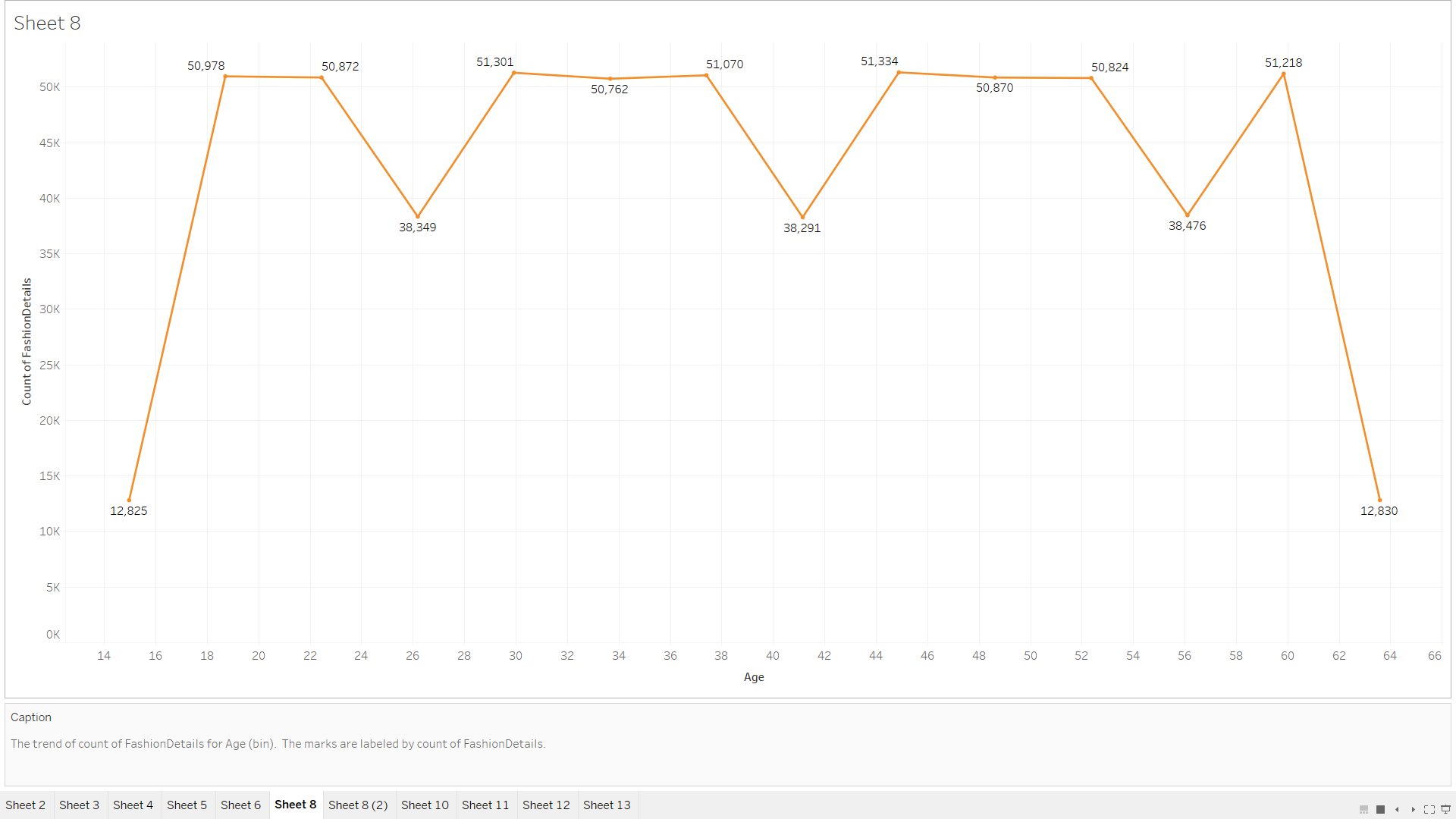


Figure 27 - The trend of count of Fashion Details for Age

* The trend of sum of Review Count for Age

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



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

# 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.

# 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.

# REFERENCES

1. **Connect Tableau**
   * [https://help.tableau.com/](https://help.tableau.com/current/pro/desktop/en-us/examples_googlebigquery.htm)
   * 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>

1. **Kaggle**

* To Download dataset
* [https://www.kaggle.com/](https://www.kaggle.com/datasets/a23bisola/fashion-dataset-uk-us)