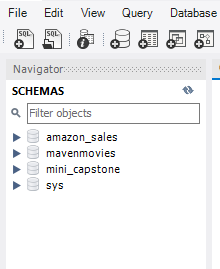
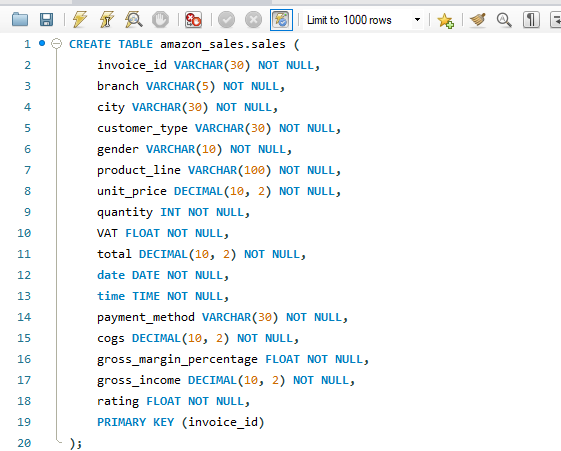
**Capstone Project – SQL**  
  
**Name: Sanjiban Hati   
Student ID: S8895**

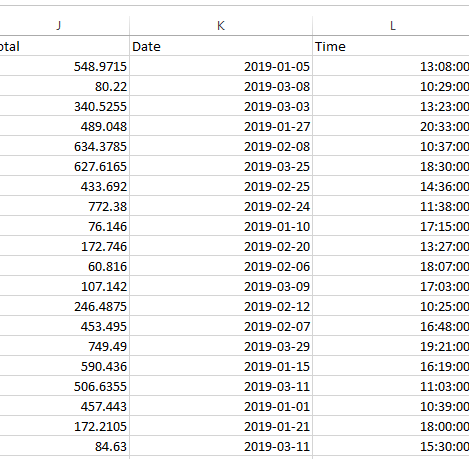
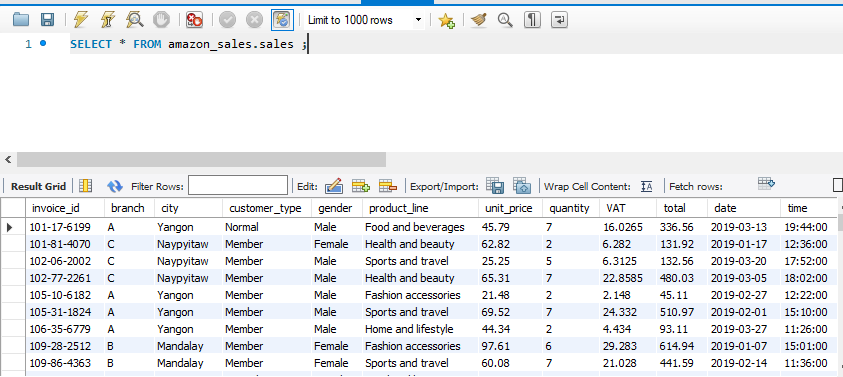
**Purposes Of The Capstone Project**

The major aim of this project is to gain insight into the sales data of Amazon to understand the different factors that affect sales of the different branches.

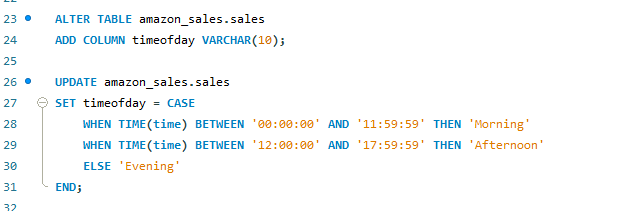
Please find below the steps that were performed as part of the execution of this project.

**Step 1: Data Wrangling**On opening MySQL Workbench and establishing a connection, the flow mentioned below was followed to create a new schema.   
**Create New Schema -> Schema Name -> Apply   
amazon\_sales** is the schema that was created for the database involved in this project.   
  
   
  
After creating the schema, was opened a new SQL tab for executing queries.   
Now, for creating the table, the following query was run.

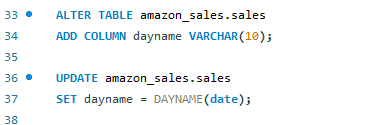


The next step was to load the data from the **Amazon.csv** file shared over Odinschool platform. But prior to that, the following steps were performed to make sure that data integrity is retained.   
  
1. First it was checked if there were any blank/null cells/rows in the table.  
2. The date format was changed from **dd-mm-yyyy** to **yyyy-mm-dd** as SQL supports the second format only.   
  
For this conversion was followed the following steps.   
i. Select the column Date  
ii. Go to Data -> Text to Columns -> Delimited -> Next -> Next (No box checked) -> Date (DMY) -> Finish  
iii. Choose the column again and right click.   
iv. Format Cells -> Custom -> yyyy-mm-dd -> OK  
  
   
  
After this the Date format was ready to be used with SQL.   
  
Once this data preparation was done, the steps mentioned below were executed for loading the data into the table which was created earlier.   
  
i. Expand the schema (amazon\_sales) and right click on the table (sales)  
ii. Choose **Table Data Import Wizard**  
iii. Browse and select the file the data needs to be loaded from   
iv. Open -> Next -> Next (Make sure the columns are correctly tagged) -> Next -> Import  
  
Once the importing gets completed, it shows a message stating **1000 rows imported**.   
  
After this, on running the following query, we are able to see the entire dataset on MySQL Workbench.   
  
  
  
As there are no null values in our database as in creating the tables, we set NOT NULL for each field while creating the table.

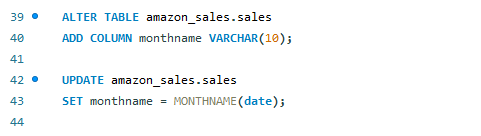
**Step 2: Feature Engineering**A new column named **timeofday** was created to give insight of sales in the Morning, Afternoon and Evening by running the following query.



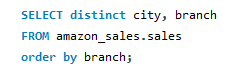
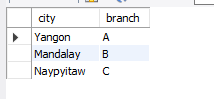
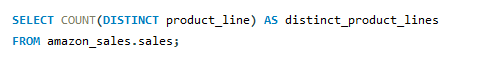
Next was created another column named **dayname** that contains the extracted days of the week on which the given transaction took place (Mon, Tue, Wed, Thur, Fri).

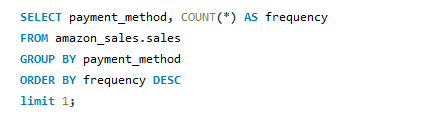


Thereafter was created a third new column named **monthname** that contains the extracted months of the year on which the given transaction took place (Jan, Feb, Mar).

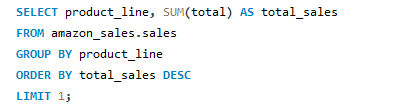


**Step 3:** **Exploratory Data Analysis (EDA)**In this segment,we ran several commands to derive conclusions as required for the business demands. The Business Questions along with the queries and outputs for the same are being recorded below for a detailed track of the analysis.

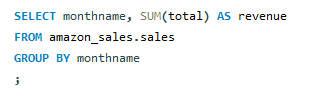
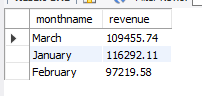
1. What is the count of distinct cities in the dataset?  
     
     
   
2. For each branch, what is the corresponding city?  
     
     
   
3. What is the count of distinct product lines in the dataset?  
     
     
   
4. Which payment method occurs most frequently?

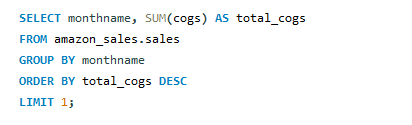

1. Which product line has the highest sales?

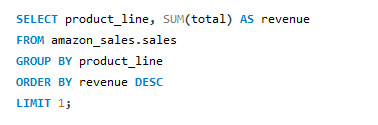

1. How much revenue is generated each month?

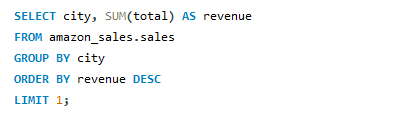
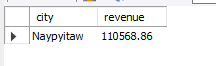
1. In which month did the cost of goods sold reach its peak?

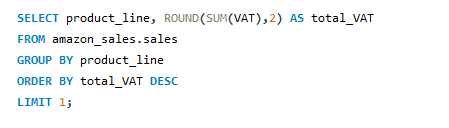
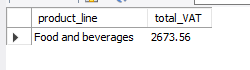

1. Which product line generated the highest revenue?

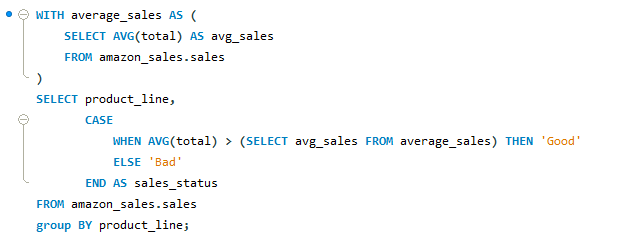

1. In which city was the highest revenue recorded?

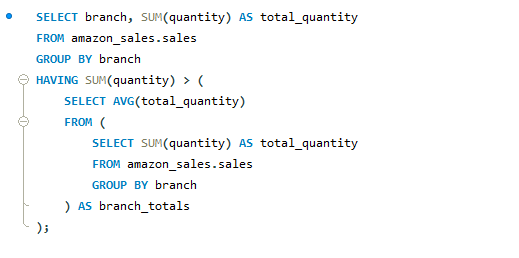
1. Which product line incurred the highest Value Added Tax?

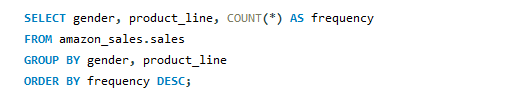
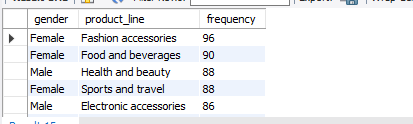
1. For each product line, add a column indicating "Good" if its sales are above average, otherwise "Bad."

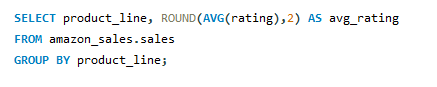
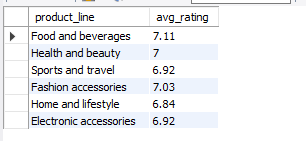

1. Identify the branch that exceeded the average number of products sold.

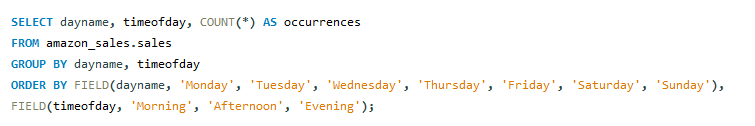
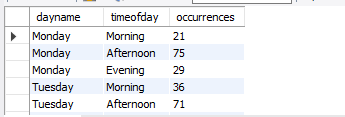

1. Which product line is most frequently associated with each gender?

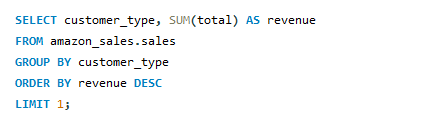
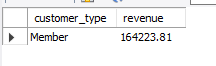
1. Calculate the average rating for each product line.

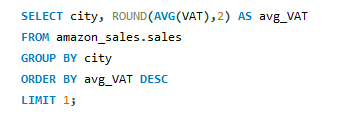
1. Count the sales occurrences for each time of day on every weekday.

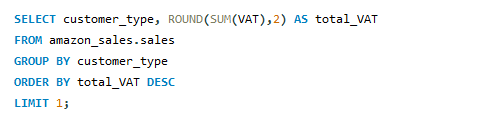
1. Identify the customer type contributing the highest revenue.

1. Determine the city with the highest VAT percentage.


1. Identify the customer type with the highest VAT payments.

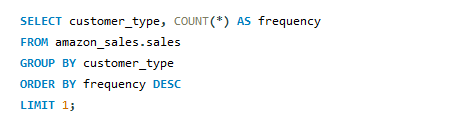

1. What is the count of distinct customer types in the dataset?

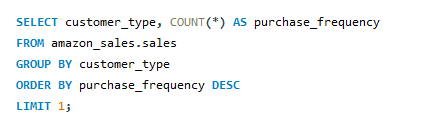

1. What is the count of distinct payment methods in the dataset?

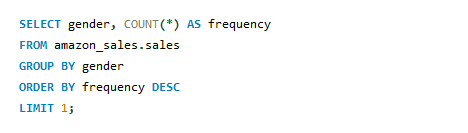

1. Which customer type occurs most frequently?

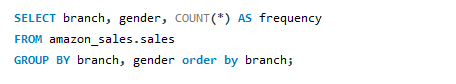
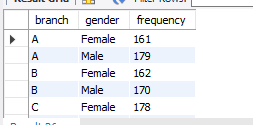

1. Identify the customer type with the highest purchase frequency.

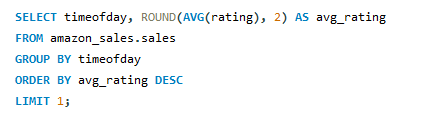

1. Determine the predominant gender among customers.

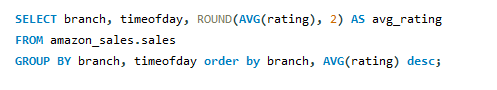
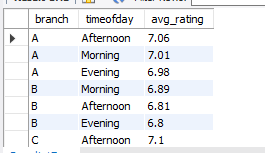

1. Examine the distribution of genders within each branch.

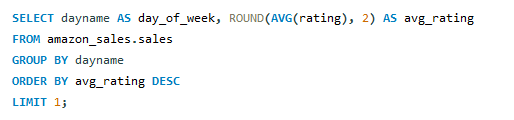
1. Identify the time of day when customers provide the most ratings.

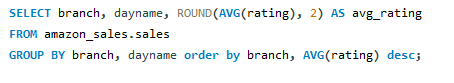
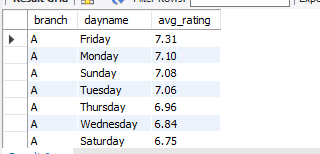

1. Determine the time of day with the highest customer ratings for each branch.

1. Identify the day of the week with the highest average ratings.


1. Determine the day of the week with the highest average ratings for each branch.

Below is being shared the list of applications which were used for executing the project:

1. MS Excel
2. MySQL Workbench (Version 8.0.33)