**Group 6**

Project Idea 6 (Outstanding): UK Train Rides

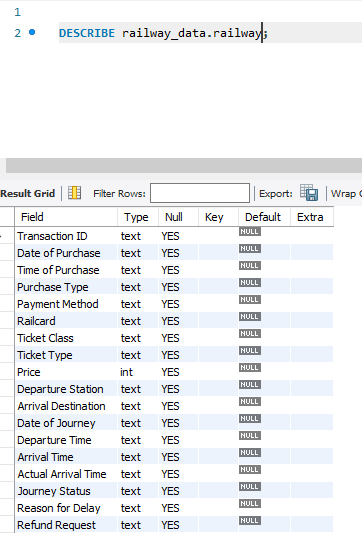
**Week 1: Build Data Model, Data Cleaning and Preprocessing**

Tasks1: Data Preprocessing:

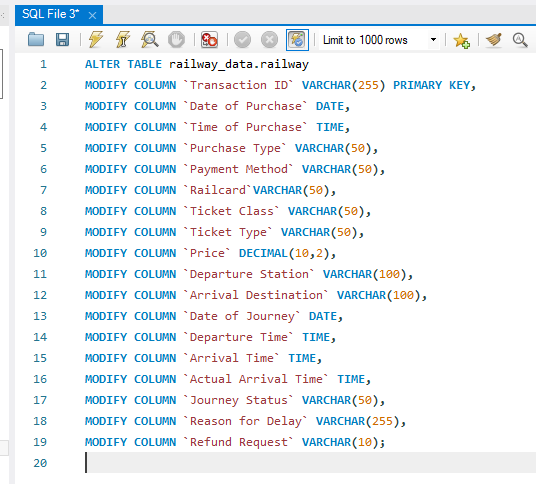
Require 1: Build a data model and clean and preprocess the data.

Tool: (SQL)

Step 1: Check Table Structure

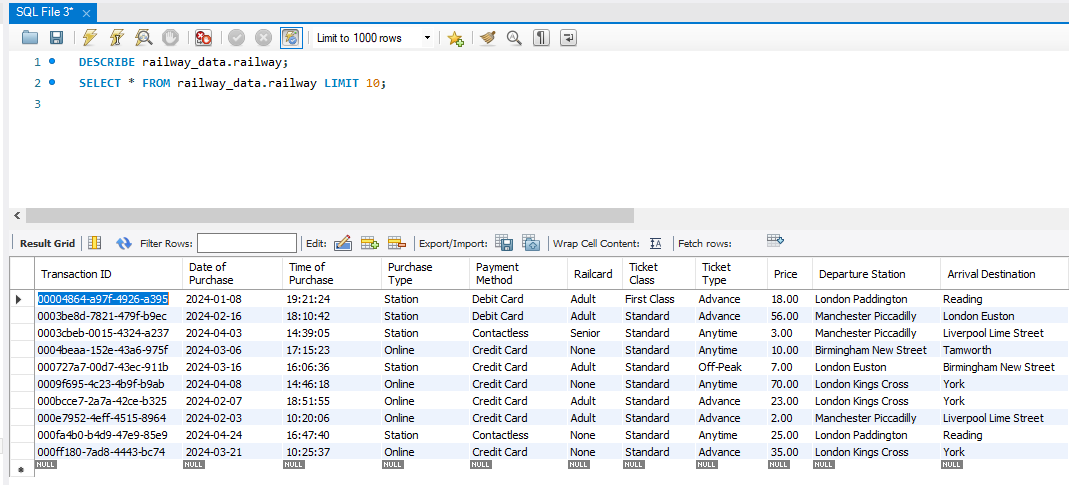


We found that: table currently has all columns set as **TEXT**, which is not optimal. We need to change the data types to match their actual usage. So, we will update data types:

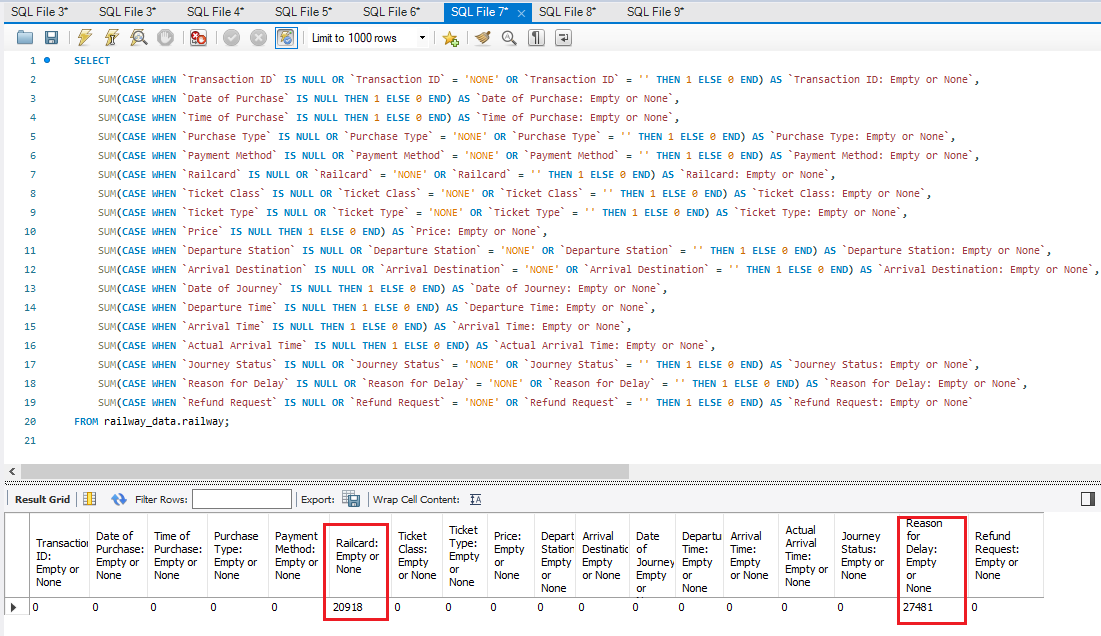


**2: Check Sample Data after updating types:**

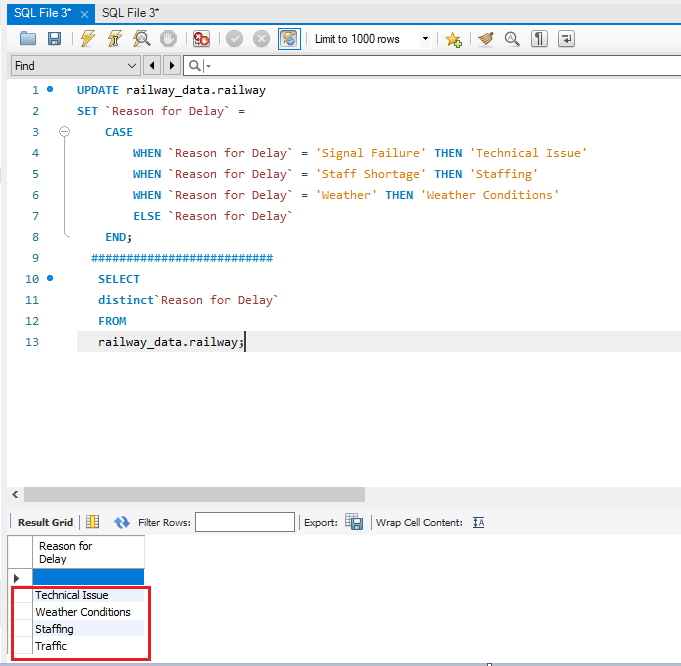
**This will give an overview of the dataset.**



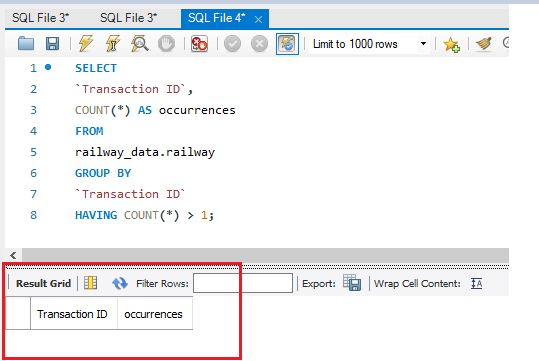
3: Check for Missing Values to helps us identify columns with missing values.



4- We identified an issue with inconsistent data entry in the "Reason for Delay" column, where similar reasons are recorded with slightly different names (e.g., "staffing" vs. "staff shortage"). This can make analysis and reporting difficult. So handled it as below:



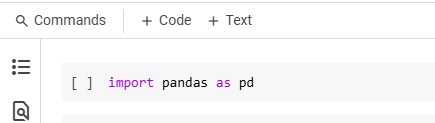
5- Checking for duplicate in Transaction ID



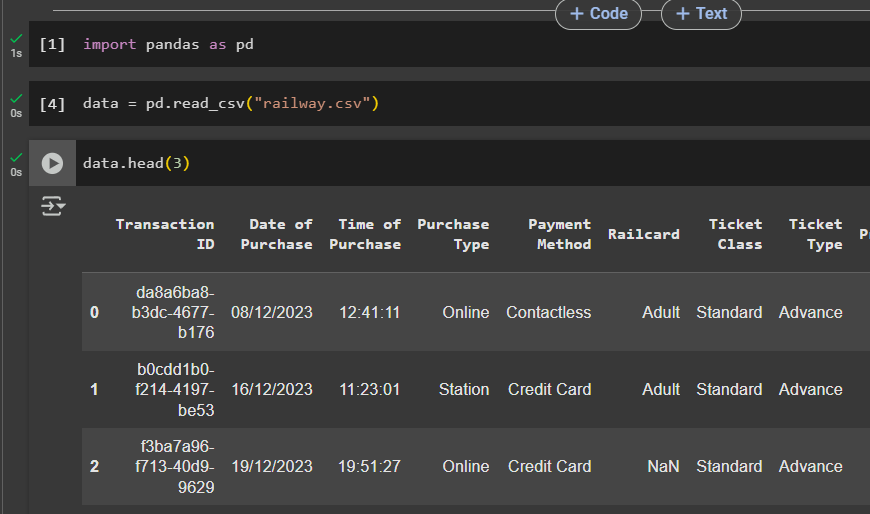
Require 2: Build a data model and clean and preprocess the data.

Tool: Python (pandas, Matplotlib)

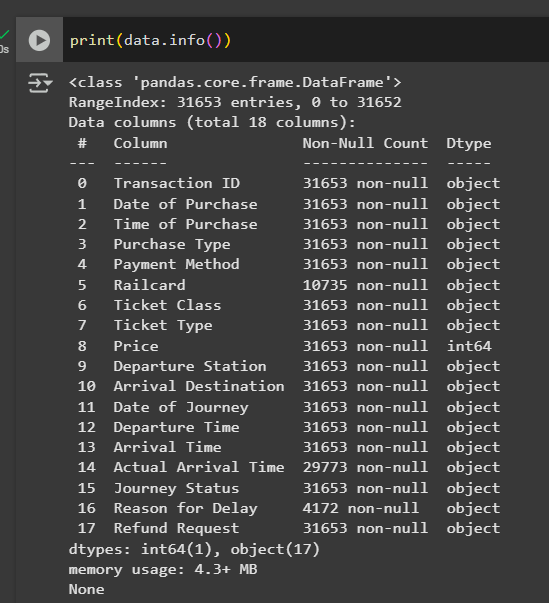
1. Import pandas



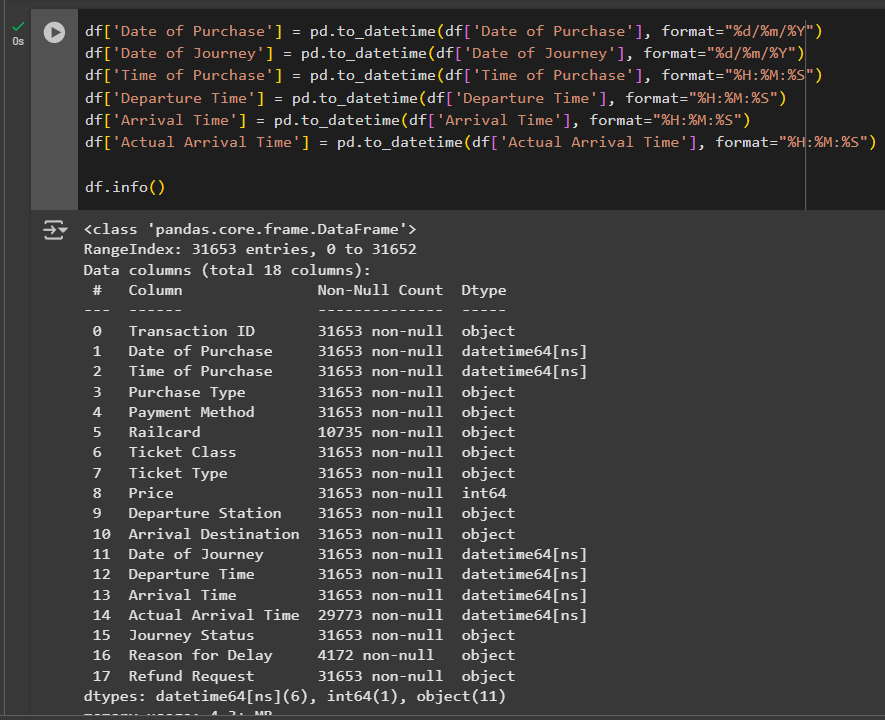
1. Upload the dataset



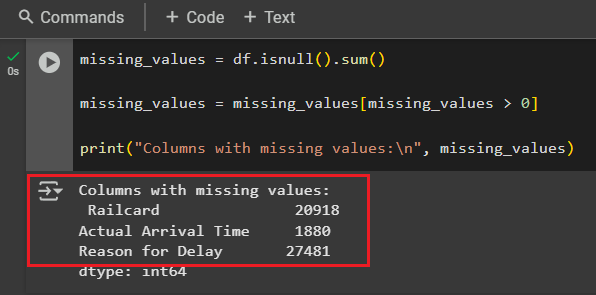
1. Check Table Structure



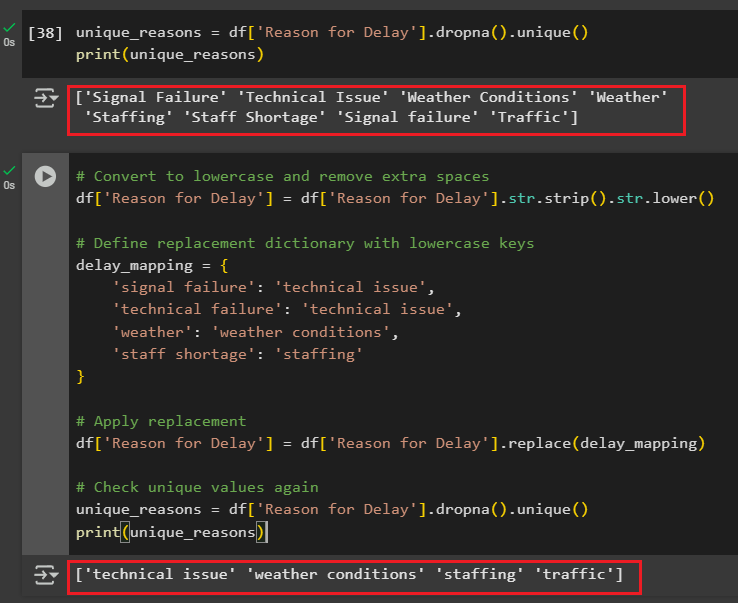
1. Amend Table Structure



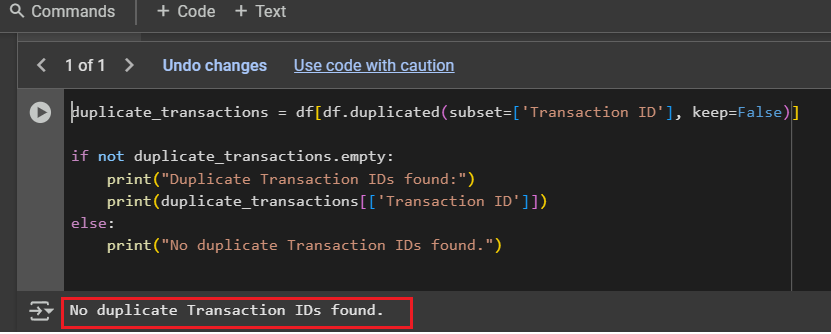
1. **Checking for missing values**



1. We identified an issue with inconsistent data entry in the "Reason for Delay" column, where similar reasons are recorded with slightly different names (e.g., "staffing" vs. "staff shortage"). This can make analysis and reporting difficult. So handled it as below:



1. 5- Checking for duplicate in Transaction ID



**Week 2: Analysis Questions Phase**

Tasks2: Determine Data Analysis Questions:

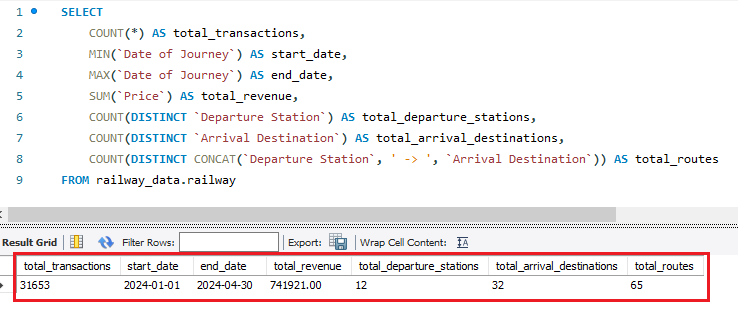
Require 1: Build a data model and clean and preprocess the data.

Tool: (SQL)

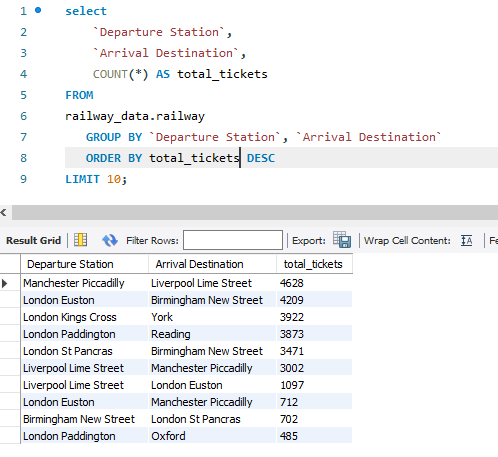
**Q1. What is total of transactions? Total revenue?**

**Q2. What is Start and end period of data?**

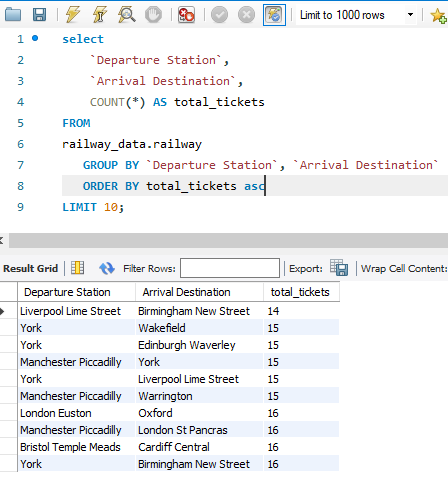
**Q3. What is total number of departure stations/arrival destinations/routes?**



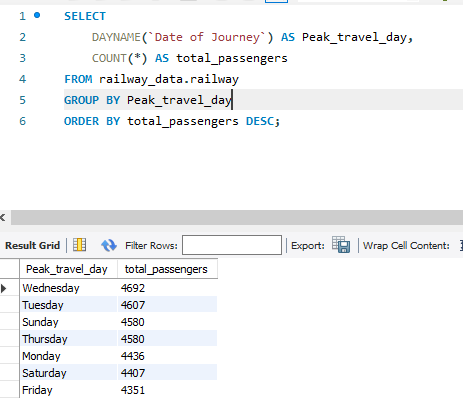
**Q4. What are the most popular train routes based on number of tickets?**



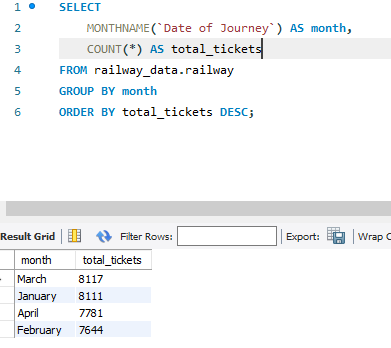
**Q5. What are the** **least popular train routes based on number of tickets?**



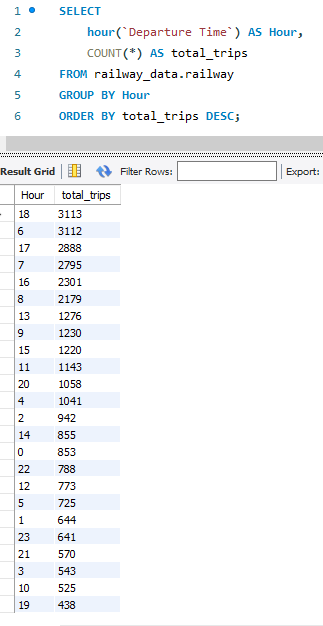
**Q6. What is the peak travel day?**



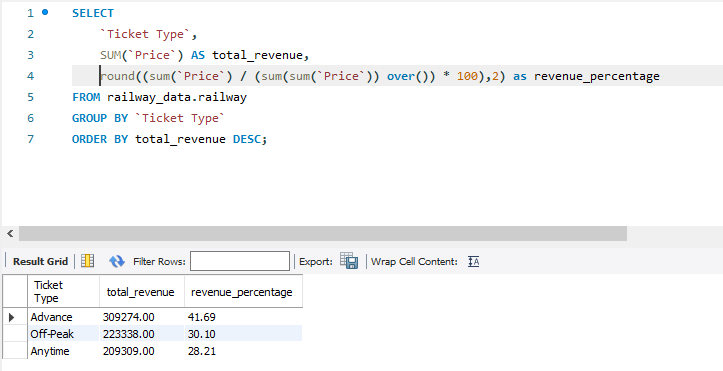
**Q7. What is the peak Month?**



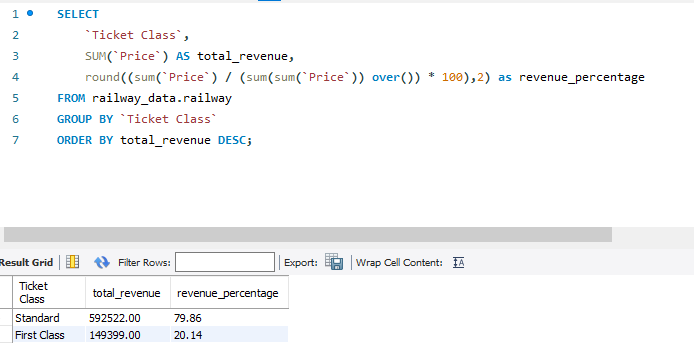
**Q8. What is the peak Hour?**



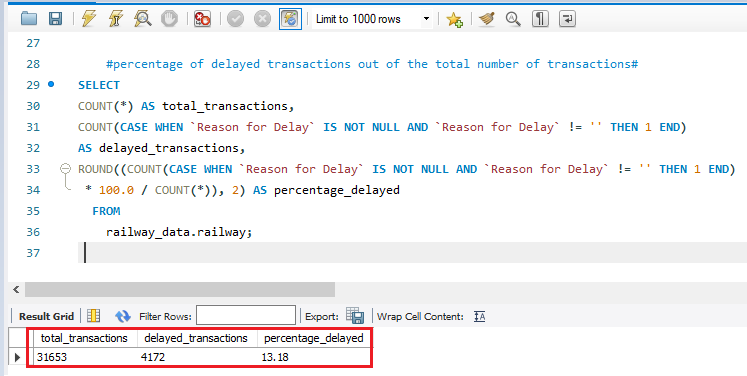
**Q.9 What is Revenue by ticket type?**



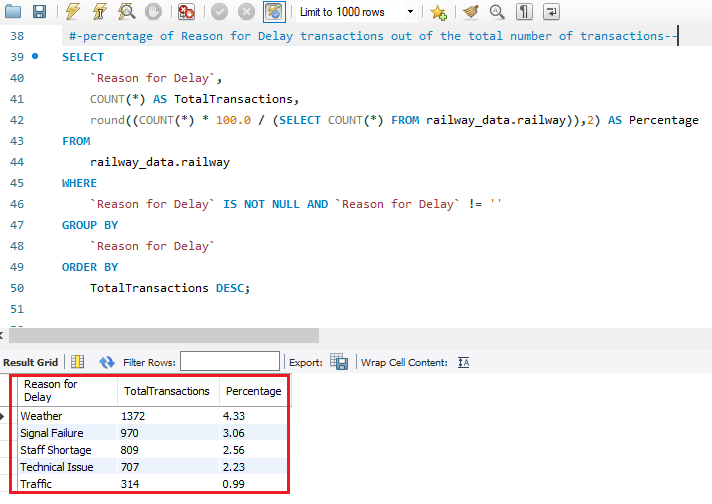
**Q.10 What is Revenue by ticket class?**



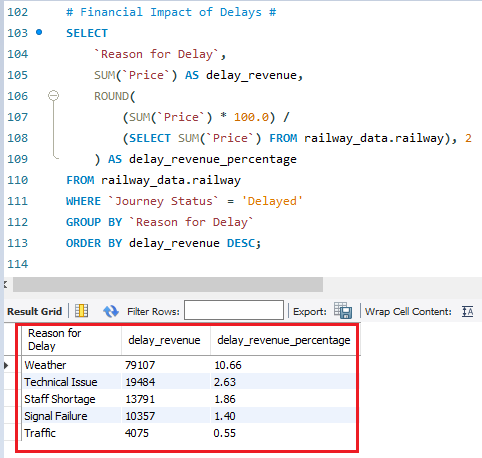
**Q.11 What is percentage of delayed transactions out of the total number of transactions?**



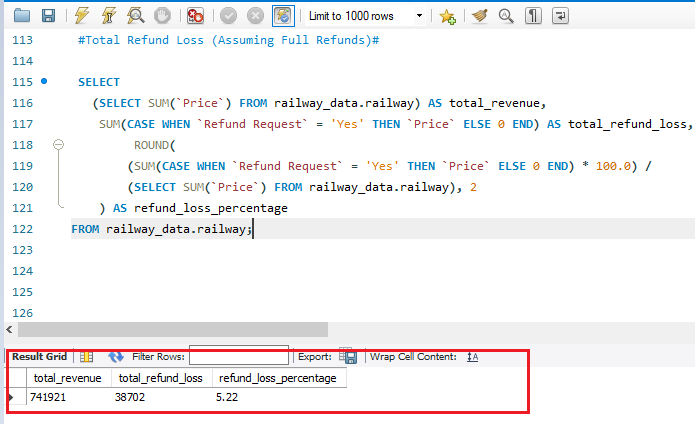
**Q.12 What is percentage of Reason for Delay transactions out of the total number of transactions?**



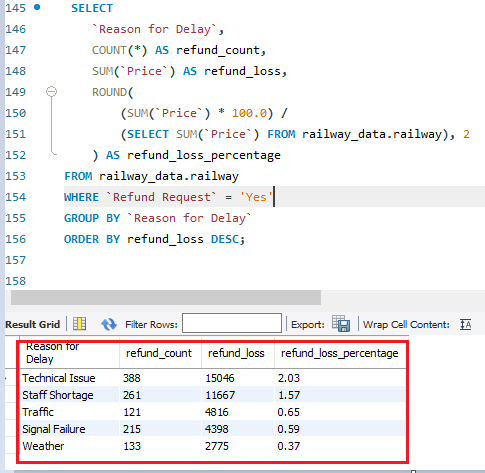
**Q.13 What is Financial Impact of Delays?**



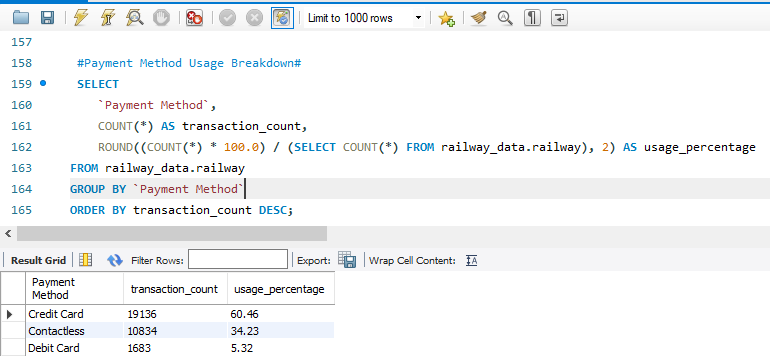
**Q.14 What is Total Refund Loss (Assuming Full Refunds)?**



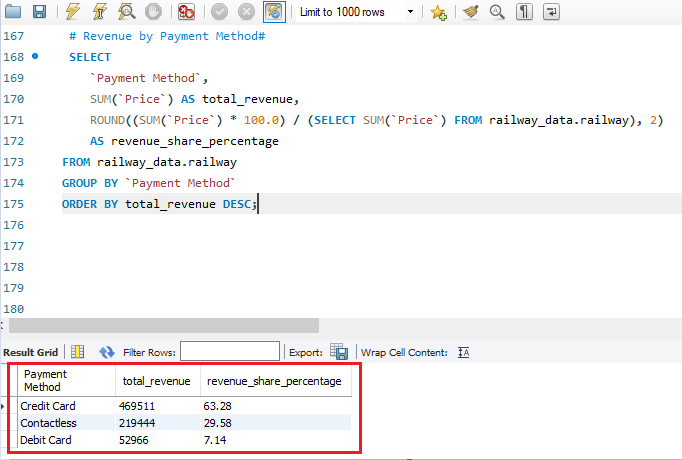
**Q.15 What is Total Refund Loss by Reason for Delay?**



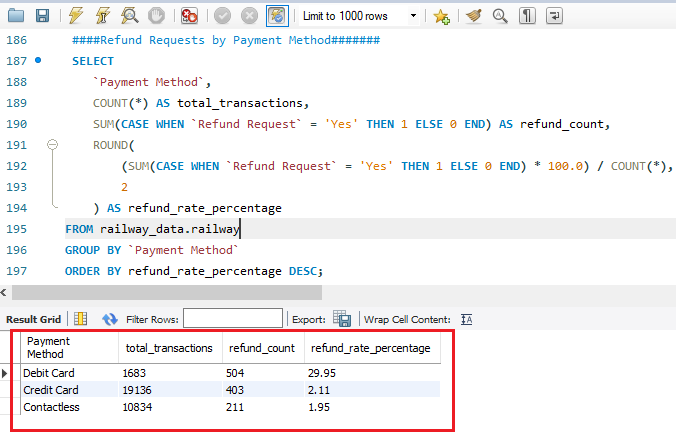
**Q.16 What is Payment Method Usage Breakdown?**



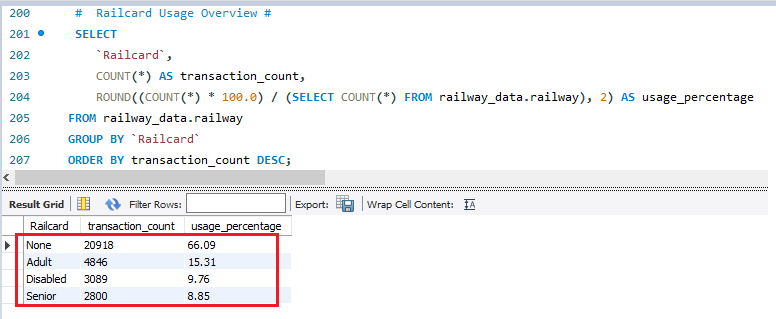
**Q.17 What is Revenue by Payment Method?**



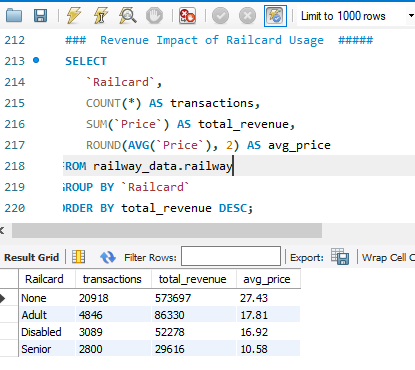
**Q.18 What is Total Refund Requests by Payment Method?**



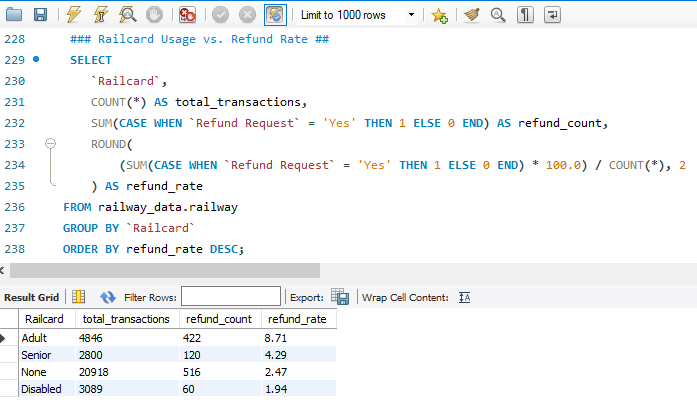
**Q.19 What is Railcard Usage Overview?**



**Q.20 What is Revenue Impact of Railcard Usage?**



**Q.21 What is Railcard Usage vs. Refund Rate?**



**Q.22 What is we Compare Railcard vs. No Railcard?**

