# 

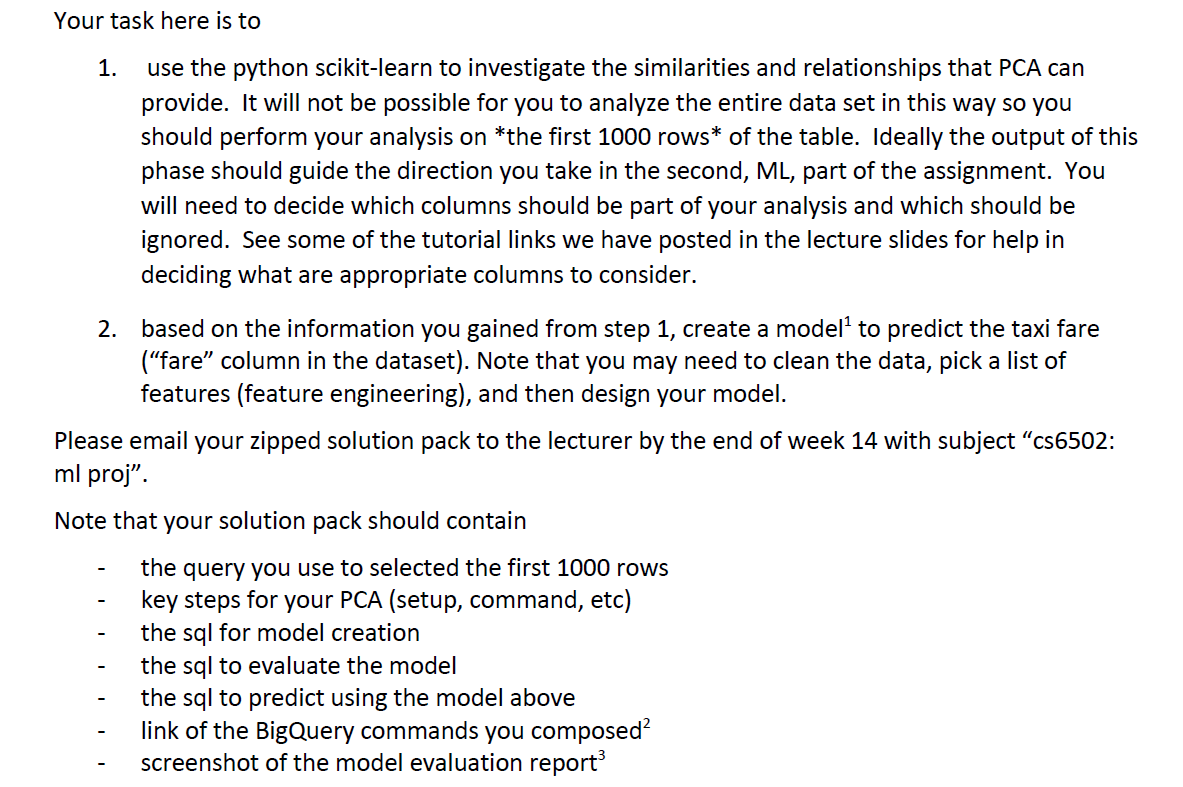
# **CS6502**

# **Applied Big Data & Visualization**

# **ML Project**

# 

# **KAILASH MURALIDHARAN 19116608**

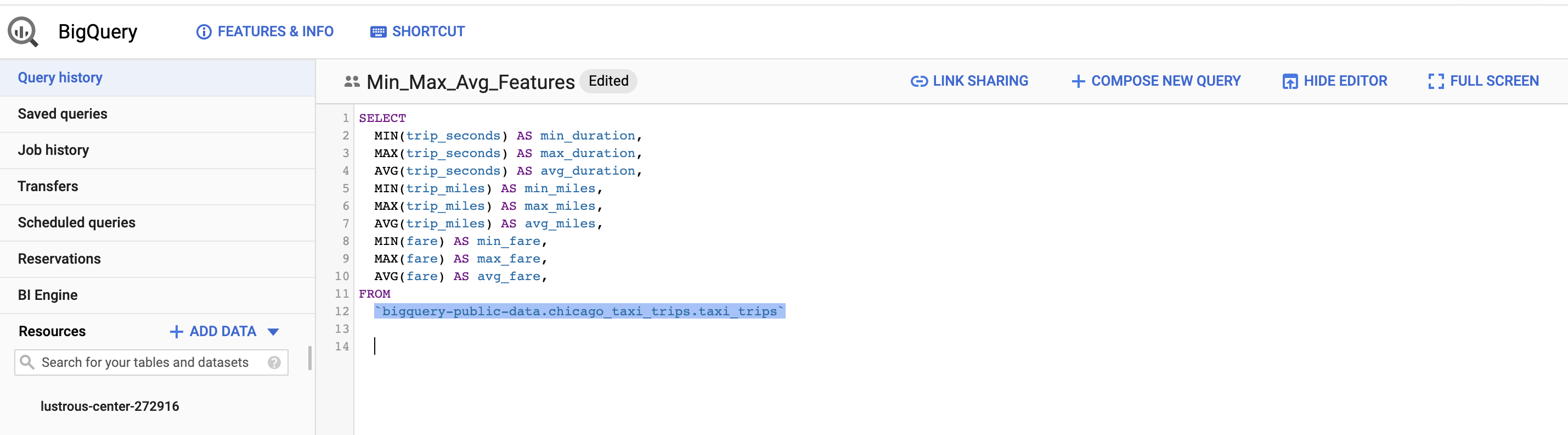


**Summary**

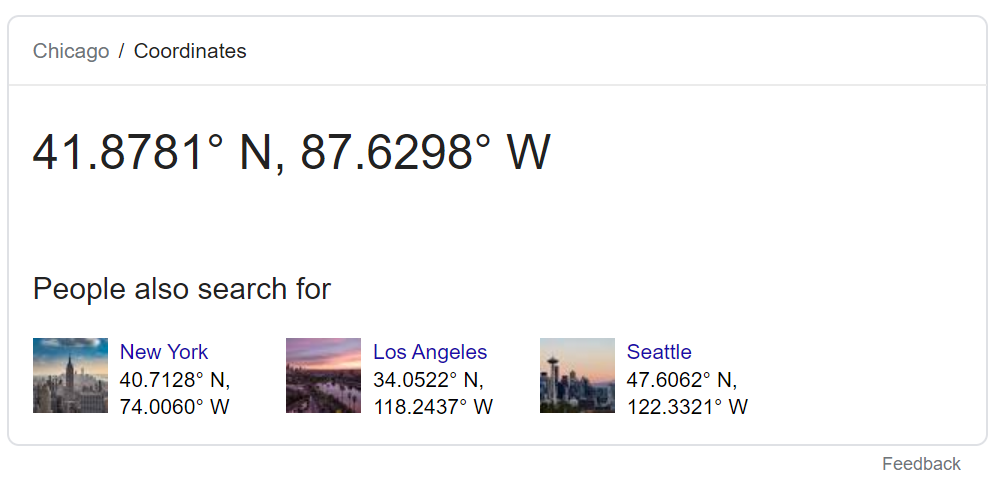
|  |  |  |
| --- | --- | --- |
|  | **Item** | **Links** |
| 1. | Query used to select 1000 rows | <https://console.cloud.google.com/bigquery?sq=3103539222:ed4b7d29551040a1a3441a2b2cf59031> |
| 2. | Results of the query containing 1000 rows | <https://docs.google.com/spreadsheets/d/1xc51HUUrw51UBCwIfbpPT343E8GyODNaQc2khHywu30/edit?ts=5eb593ba#gid=2010390566> |
| 3. | Query used to create the Final Model | <https://console.cloud.google.com/bigquery?sq=3103539222:88ef2a11074c451db05d9e1598f90d55> |
| 4. | Query used for Model Evaluation | <https://console.cloud.google.com/bigquery?sq=3103539222:4c44667fd77d489e8a8164b3a02b843c> |
| 5. | Query used for Model Prediction | <https://console.cloud.google.com/bigquery?sq=3103539222:4be462936e3c40eba4f80b02cc69c4b4> |
| 6. | Final Model Evaluation Report |  |

**1. Query to select 1000 rows**

I have used a query to check for the minimum, maximum and average of three numeric columns.

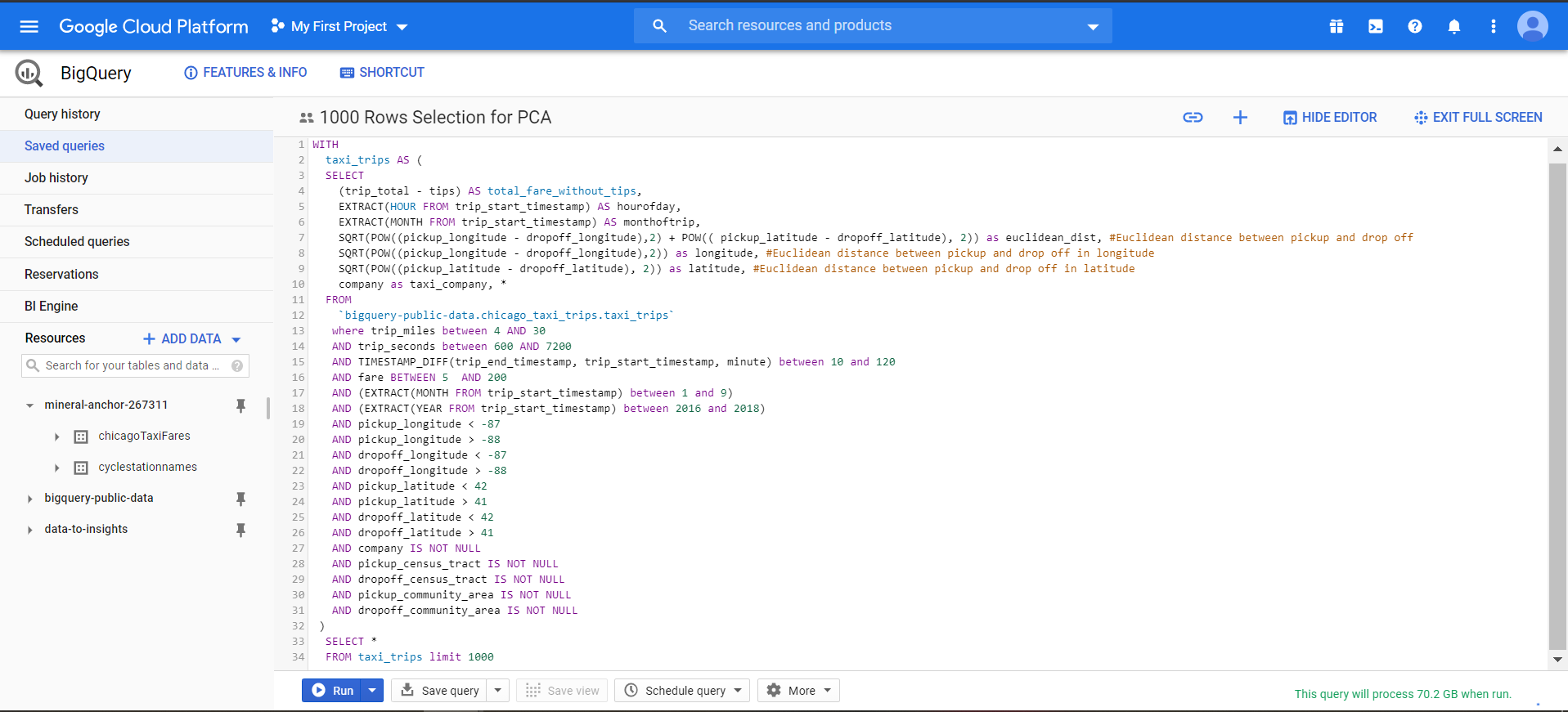


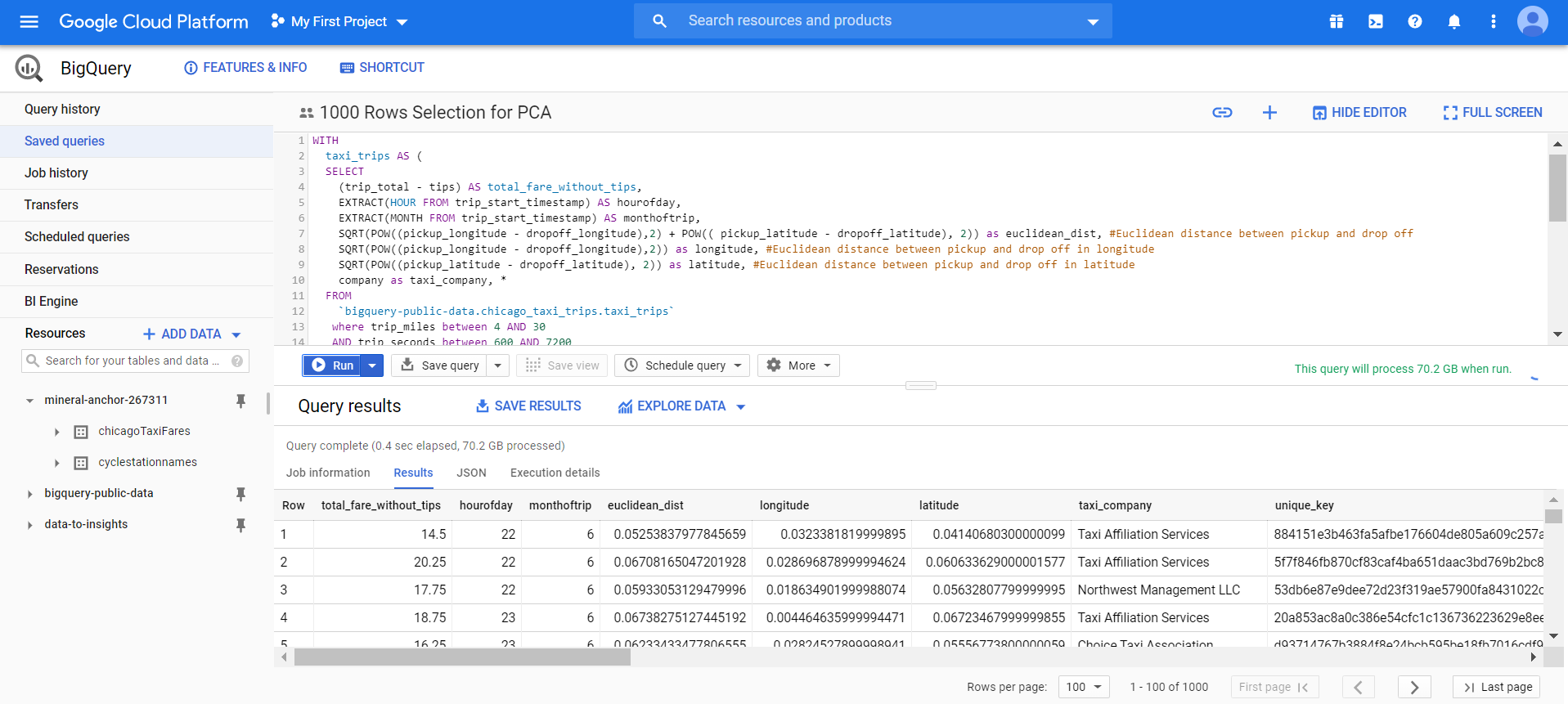
Using the latitude and longitude values of Chicago, to consider only those trips that started and ended inside the city range.

****

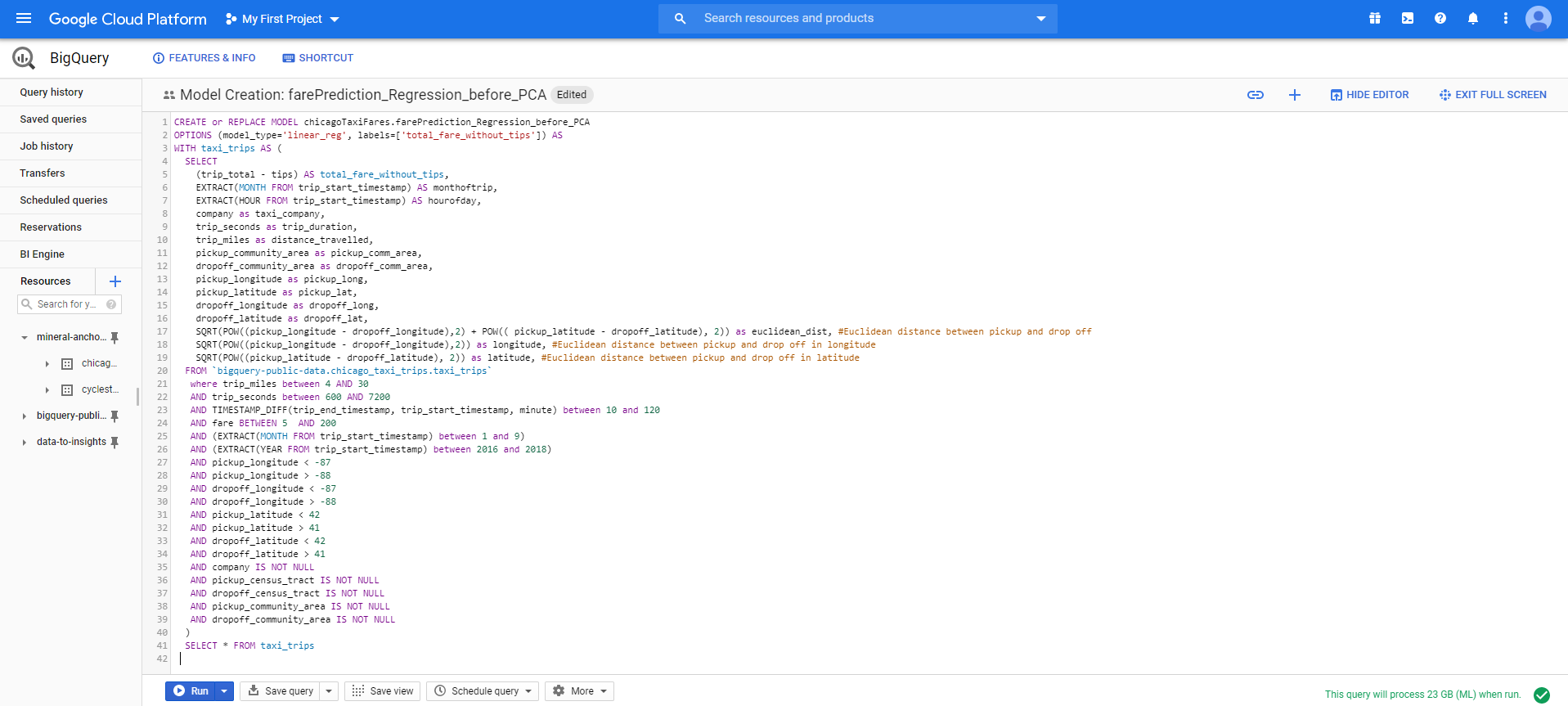
**Conditions applied:**

* Picked values to apply for where conditions based on the min, max, avg summary and Chicago’s location coordinates.
* Years filtered between 2016 and 2018, and months between January and September.
* Rest of the columns – NOT NULL condition
* In addition to the existing columns, following additional columns have been created
* **total\_fare\_without\_tips** (the value to be predicted), since the tips amount varies from customer to customer (created a column by subtracting tips from the trip\_total)
* **hour\_of\_day**, from the timestamp
* **month\_of\_trip**, also from the timestamp
* **euclidean\_distance**, the distance between the pickup and dropoff points
* **longitude,** the distance between the pickup and dropoff points longitude
* **latitude,** the distance between the pickup and dropoff points latitude
* **taxi\_company**, the name of the taxi company

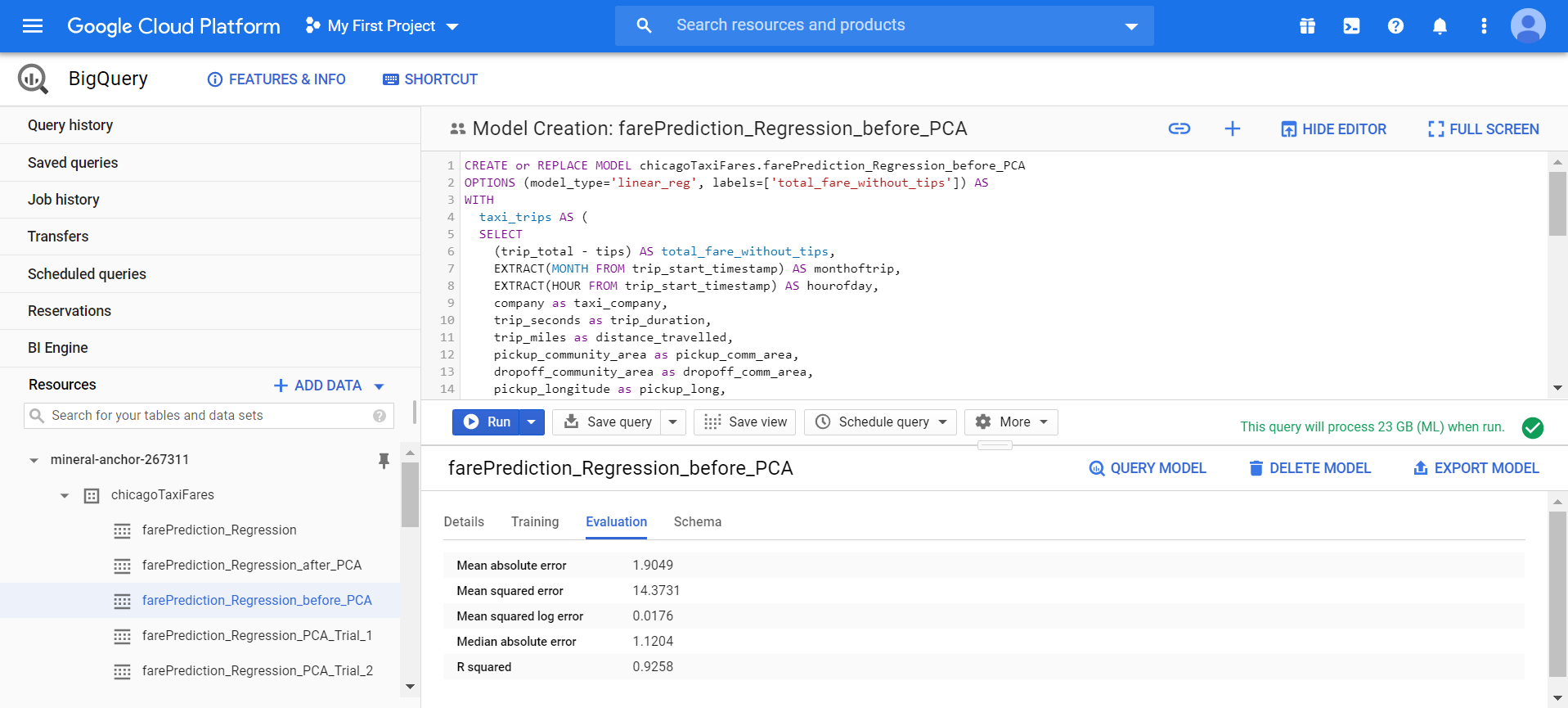




**2. Base Model before PCA**



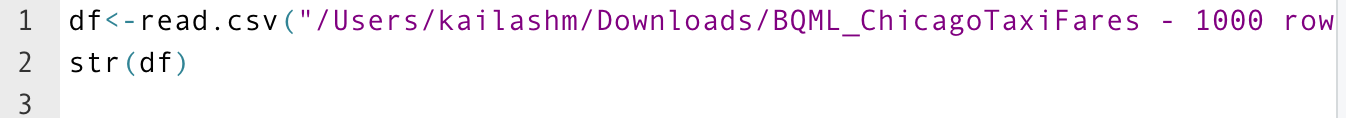
**Evaluation Report:**

****

* MSE= **14.3731**
* **R squared value** = **0.9258**, about **92.58**% of the variability in the dependent variable is explained by our model.

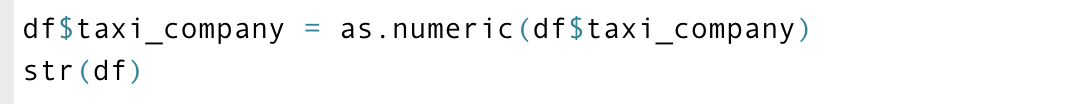
**3. PCA using R statistical package**

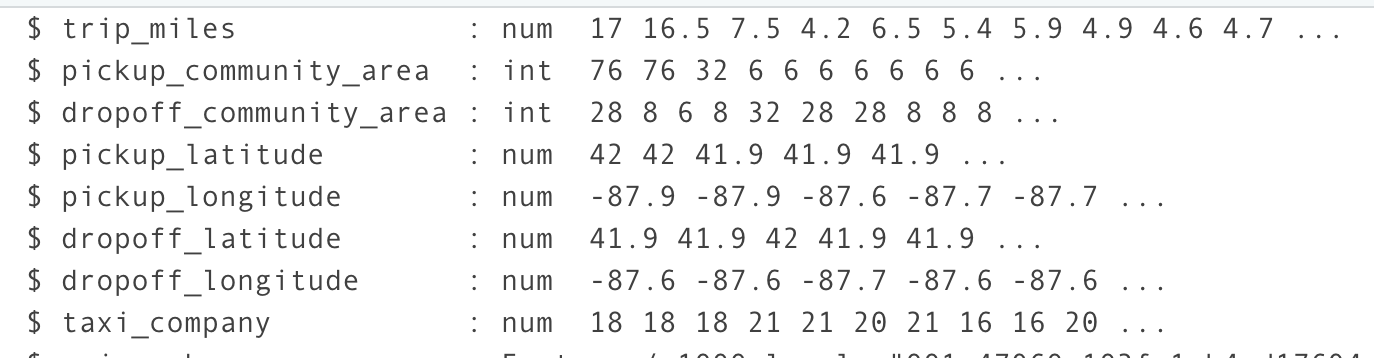
* Load the dataset into the dataframe df



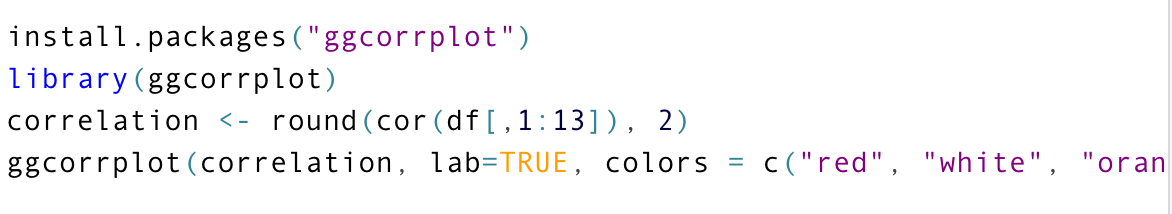
****

* The taxi\_company column is changed to be of numeric data type (to be included in creating the model)

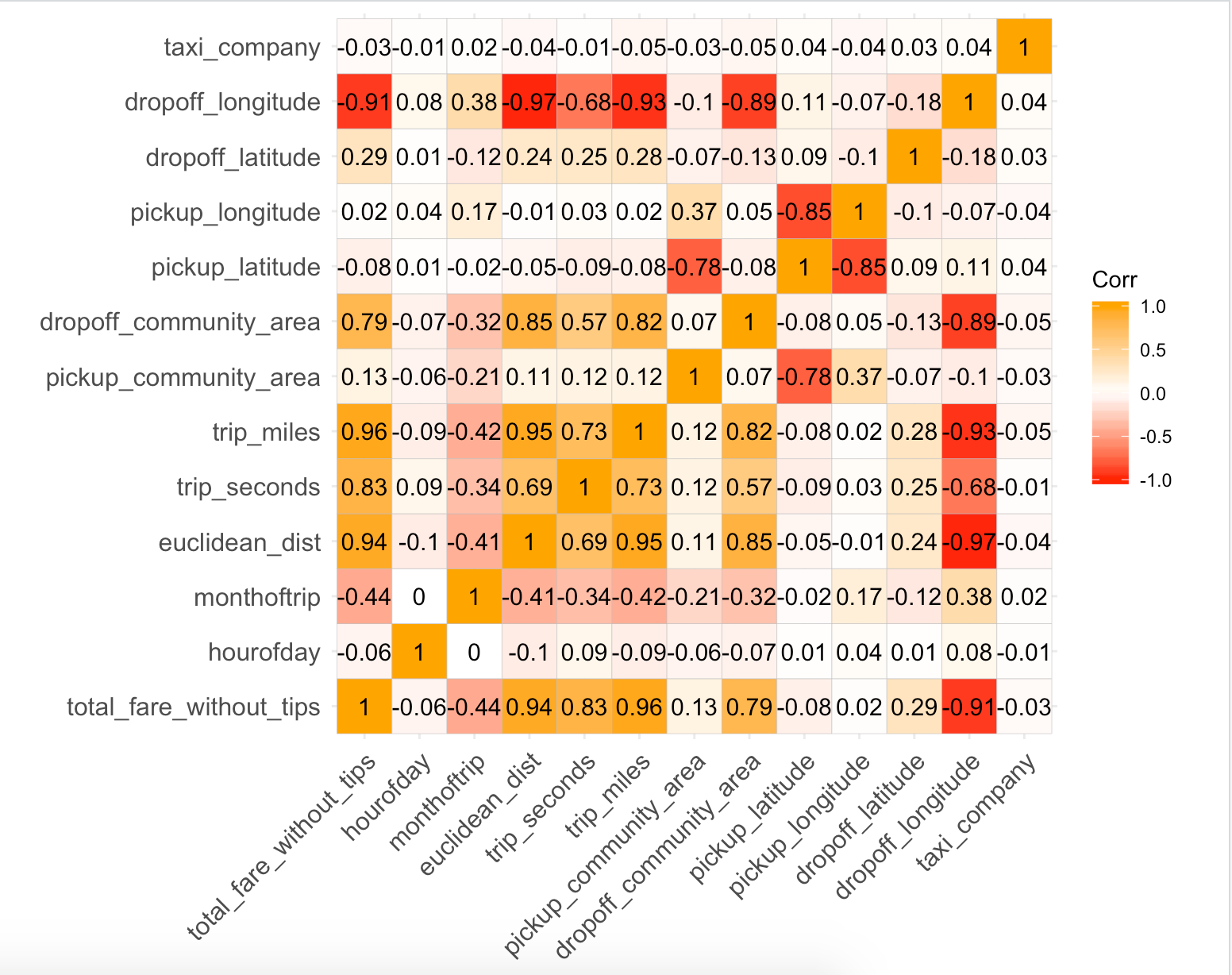
****

****

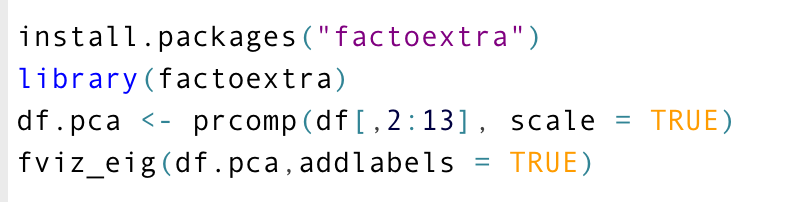
* Correlation Matrix

****

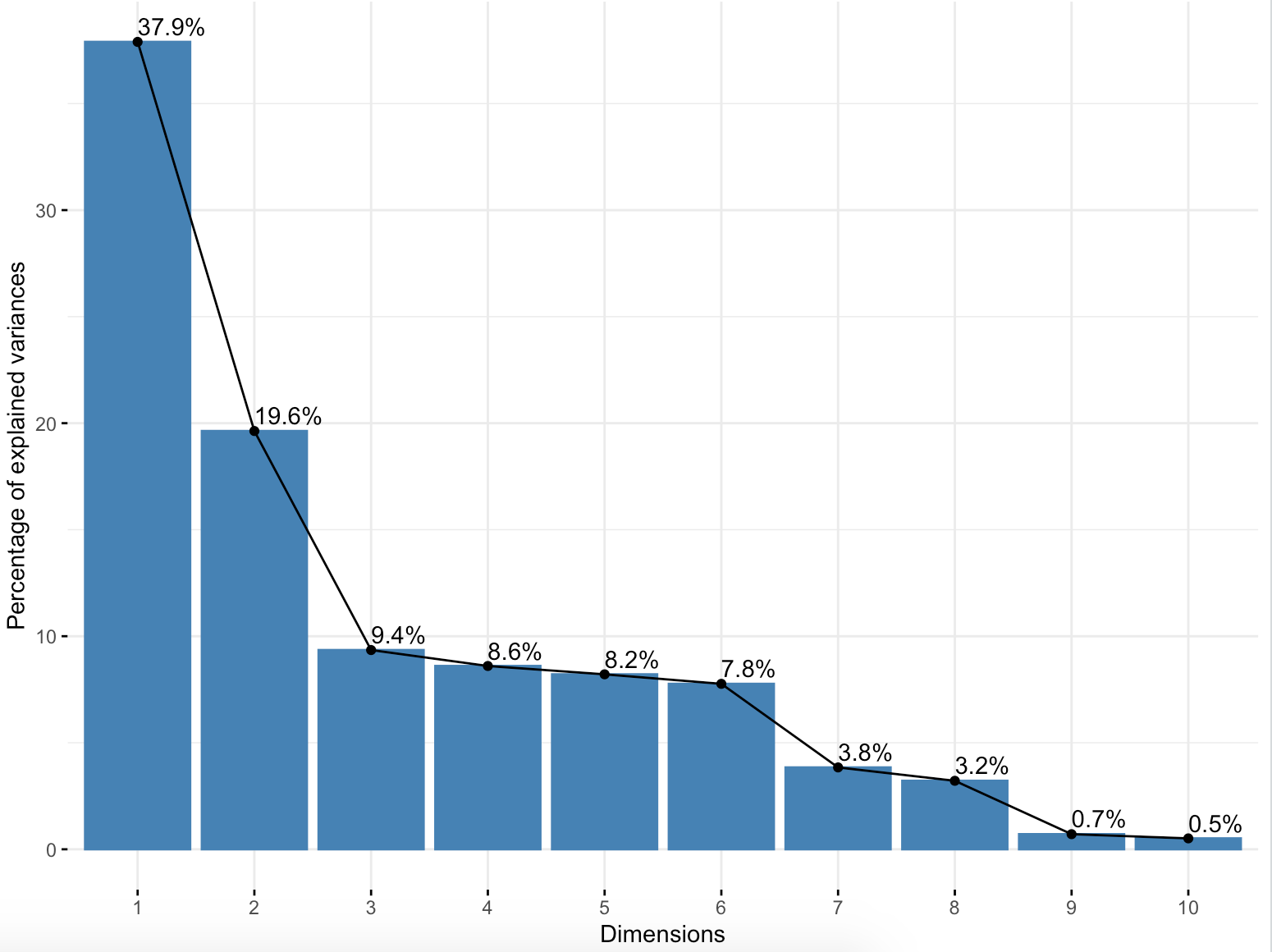
The below matrix shows the correlation between the numeric variables in the dataset.

****

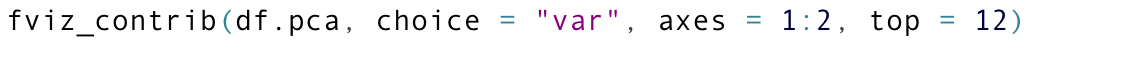
* Computing PCA using prcomp()

****

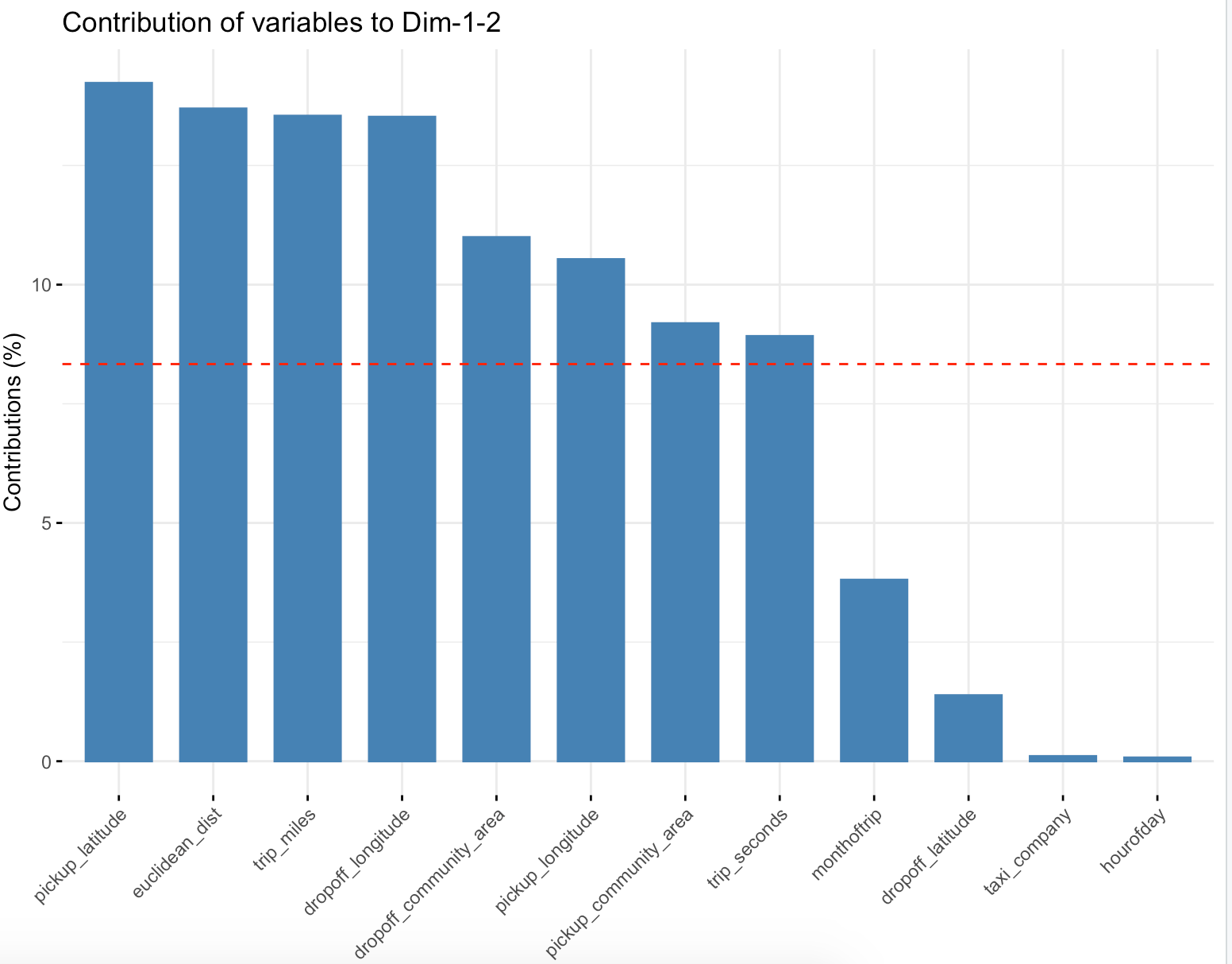
**This plot shows the percent contribution of the principle components.** We can see that the first two principle components explain about **57.5%** variation.

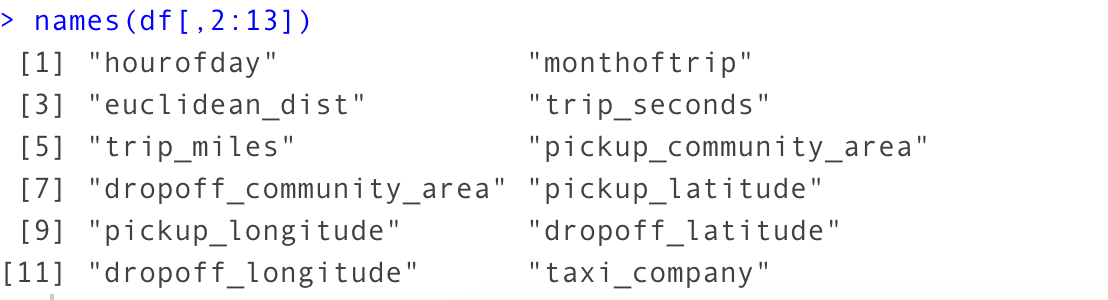
****

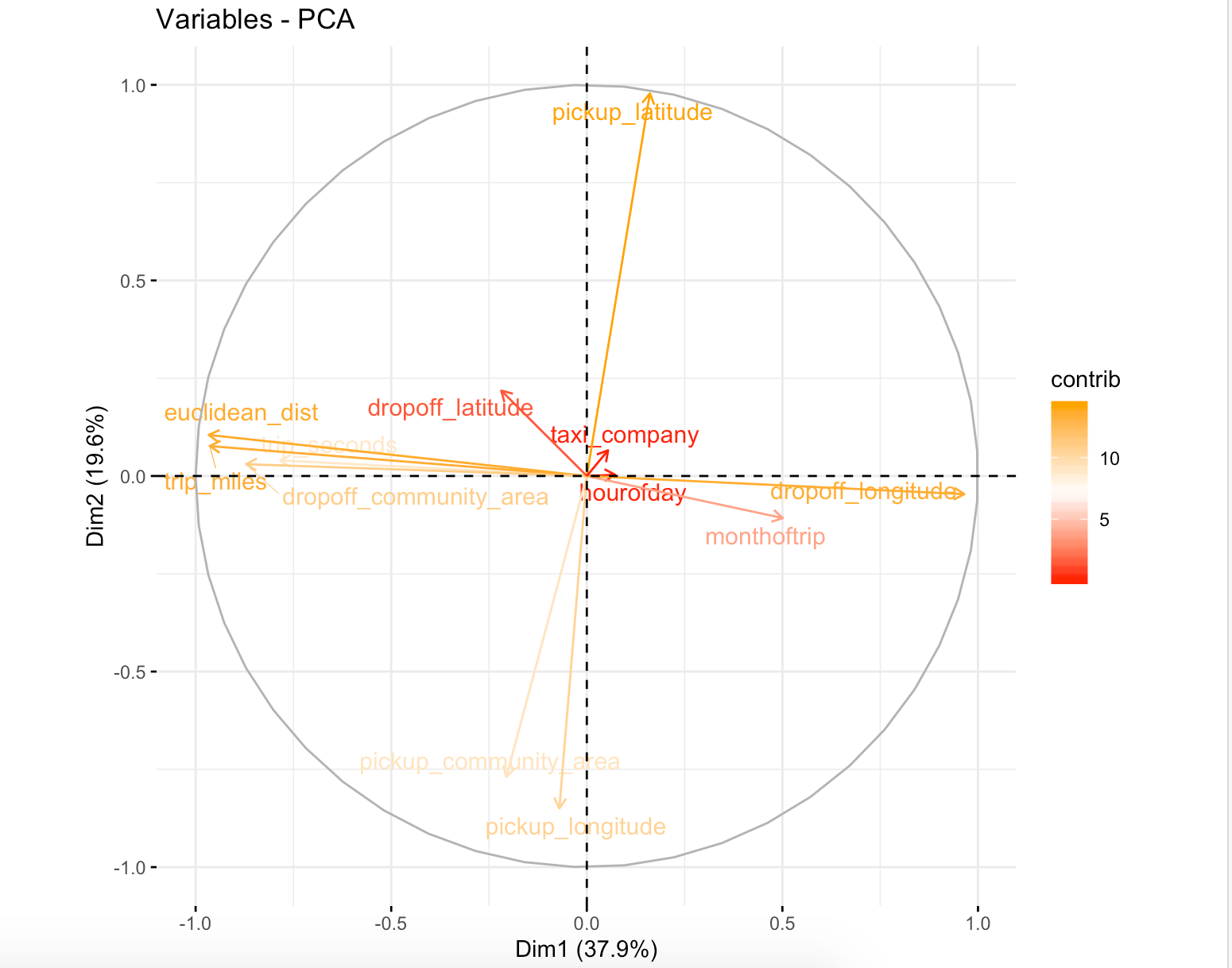
* We visualize the first two principle components to check the **% contribution of each variable** (features from the original dataset) **in both the principle components**.

****

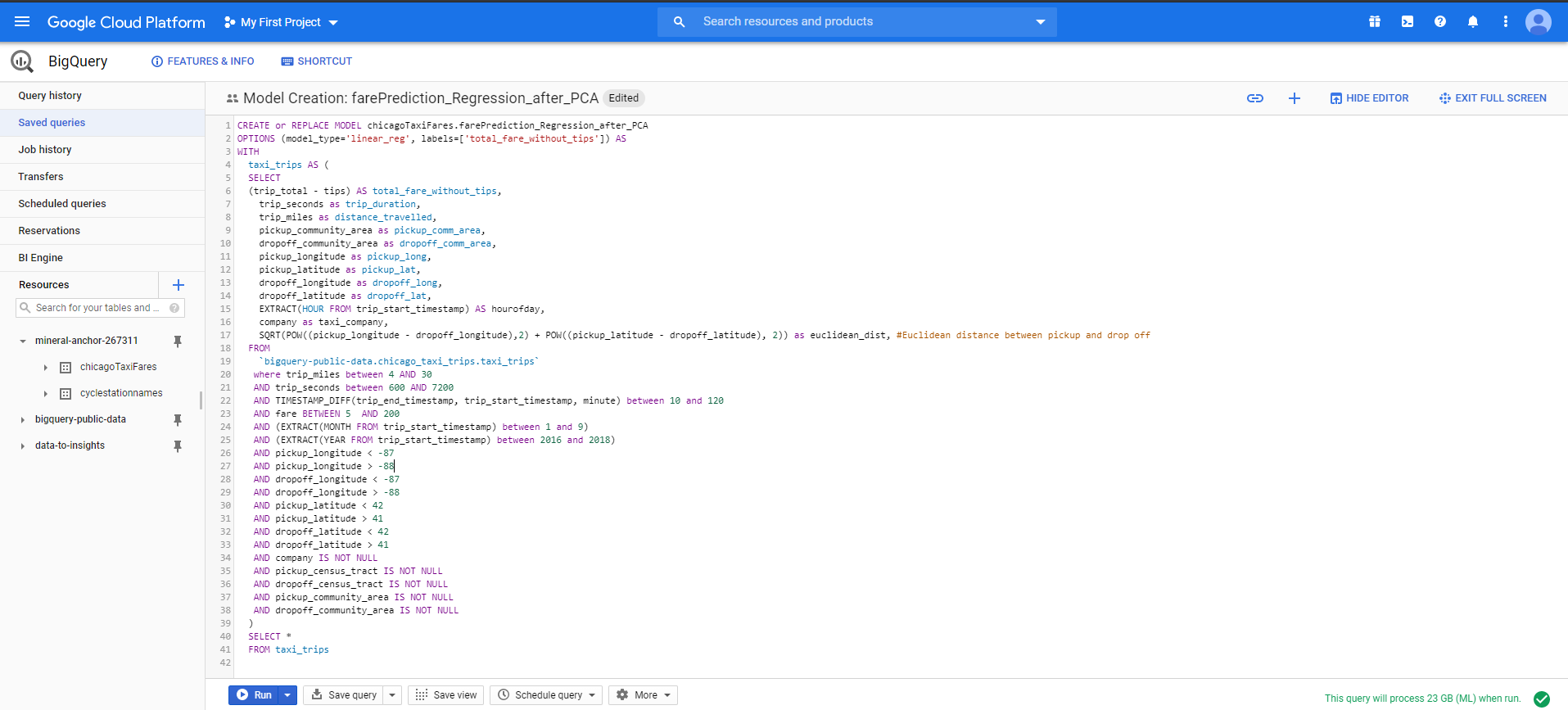
**Output Screenshot:**

****

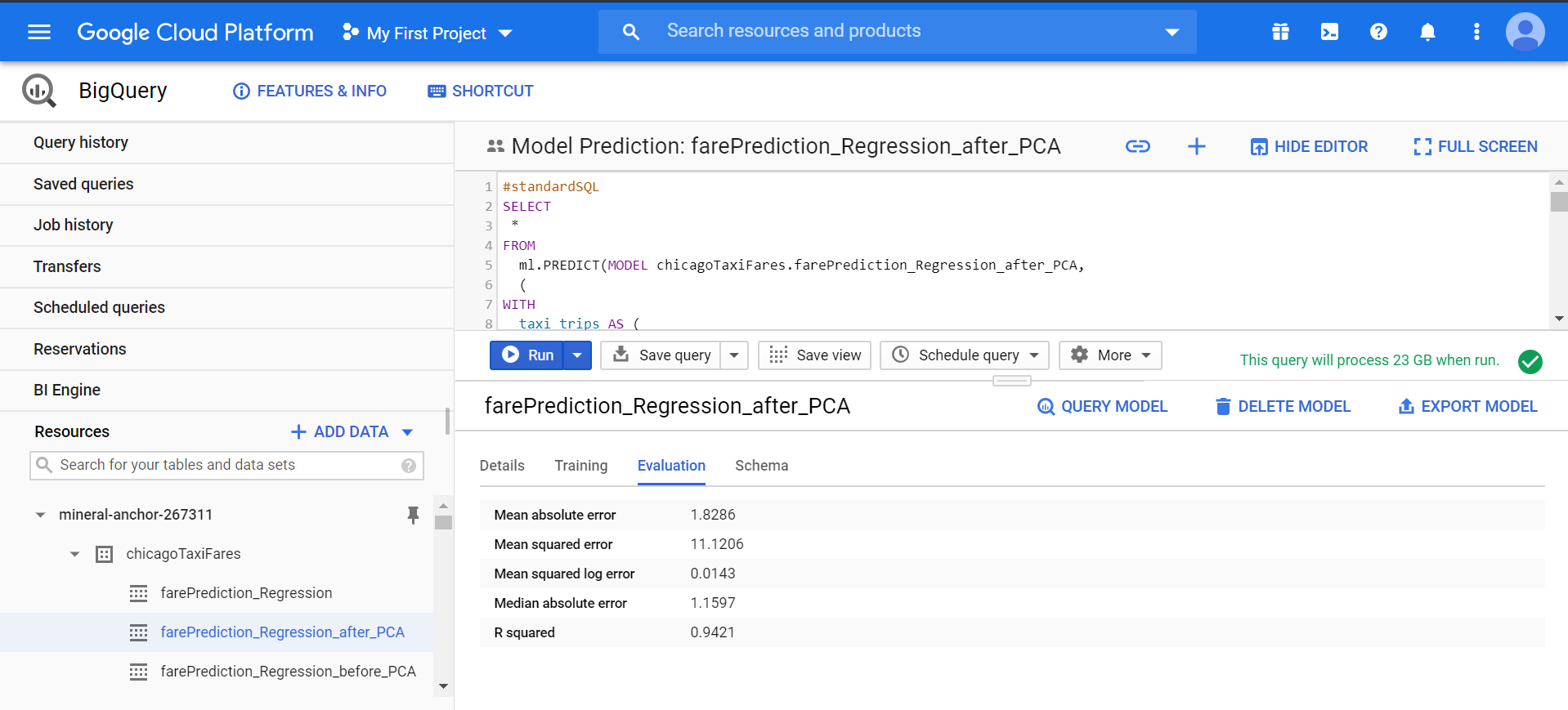
****

****

**4. Final Model after PCA and feature selection**

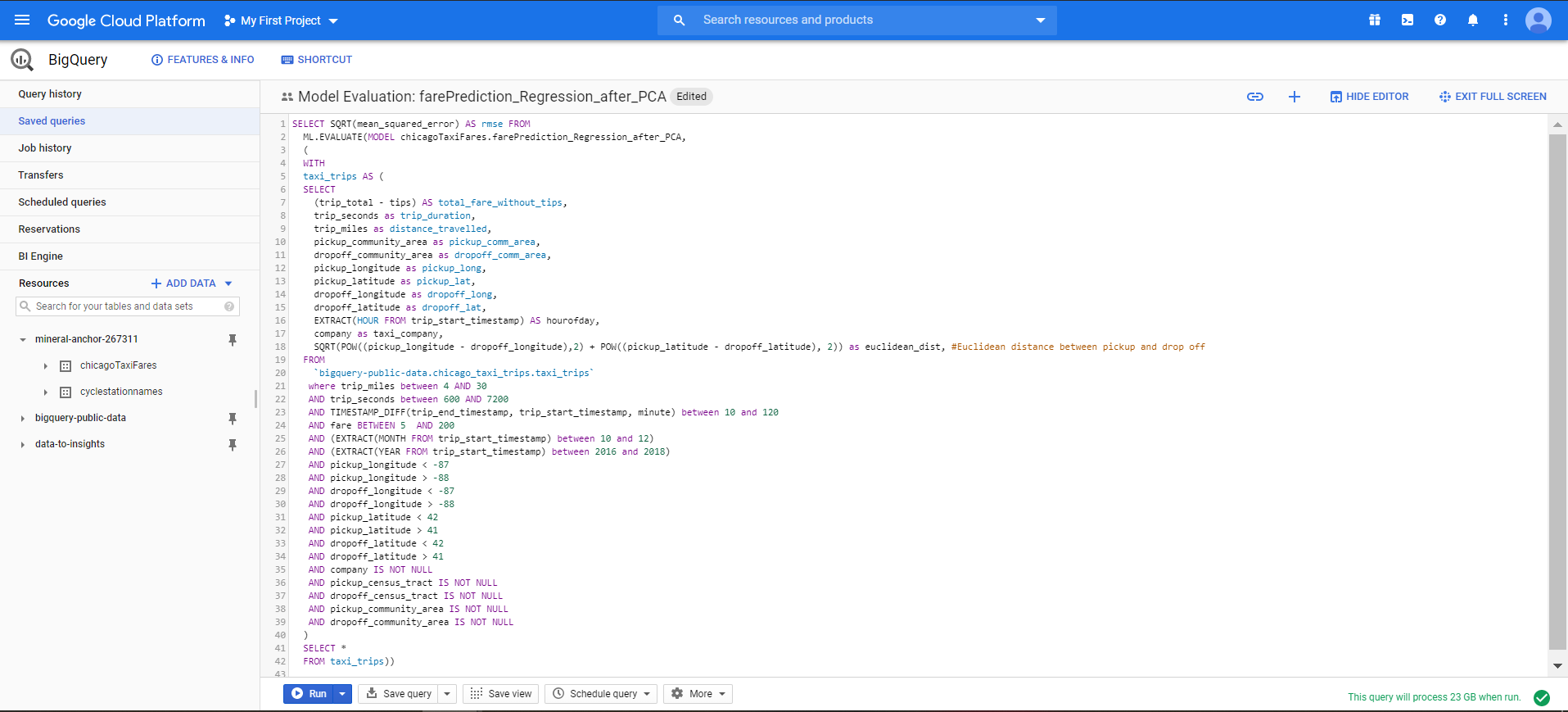
****

**Final Prediction Model - Evaluation Report:**

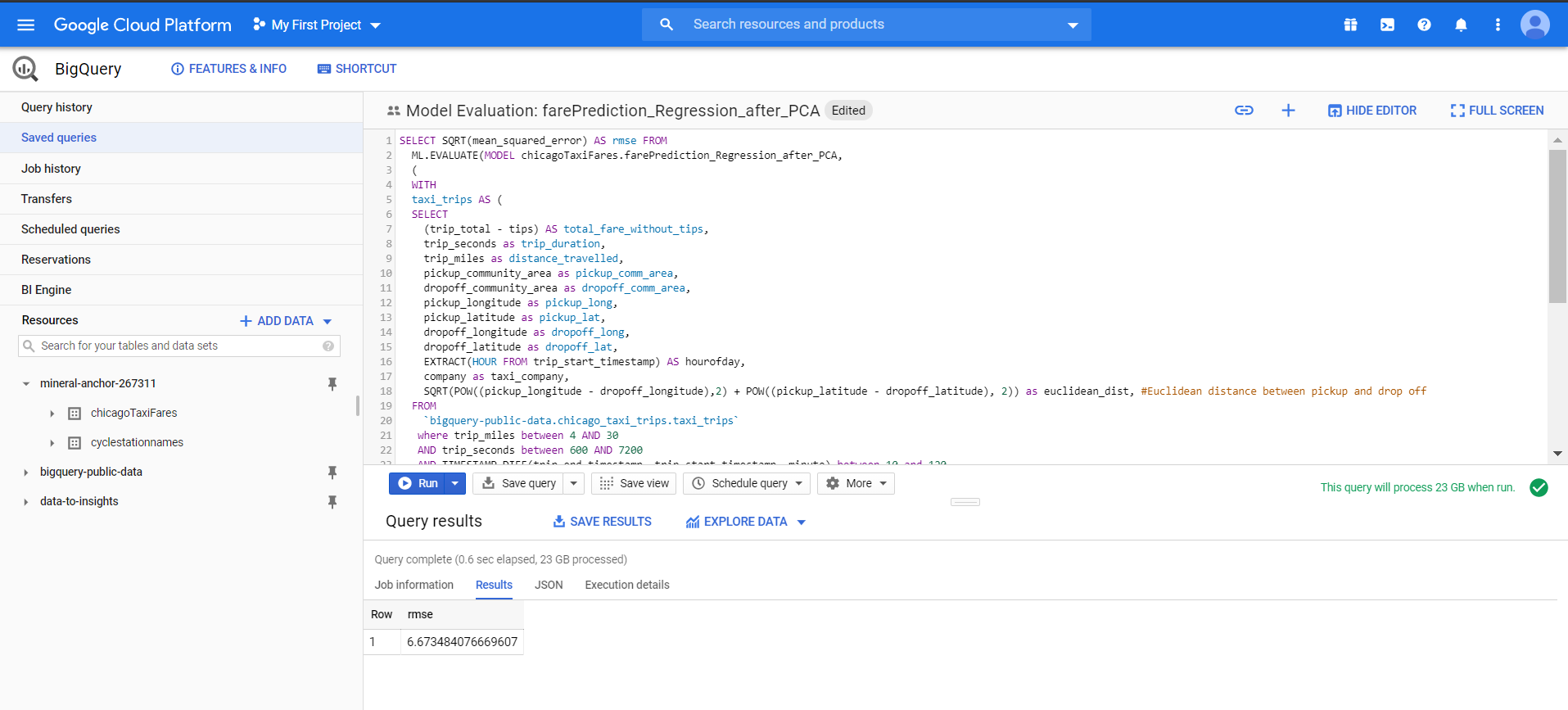
****

* The MSE of the prediction model **after PCA,** is **11.1206**, **which** **is** **less than** that of the **base model MSE = 14.3731**
* The **R squared value** of the prediction model has **increased** from **0.9258** to **0.9421**, which suggests that the final model captures more variation in the **total\_fare\_without\_tips** than the base model

**5. Final Model - Evaluation Query**

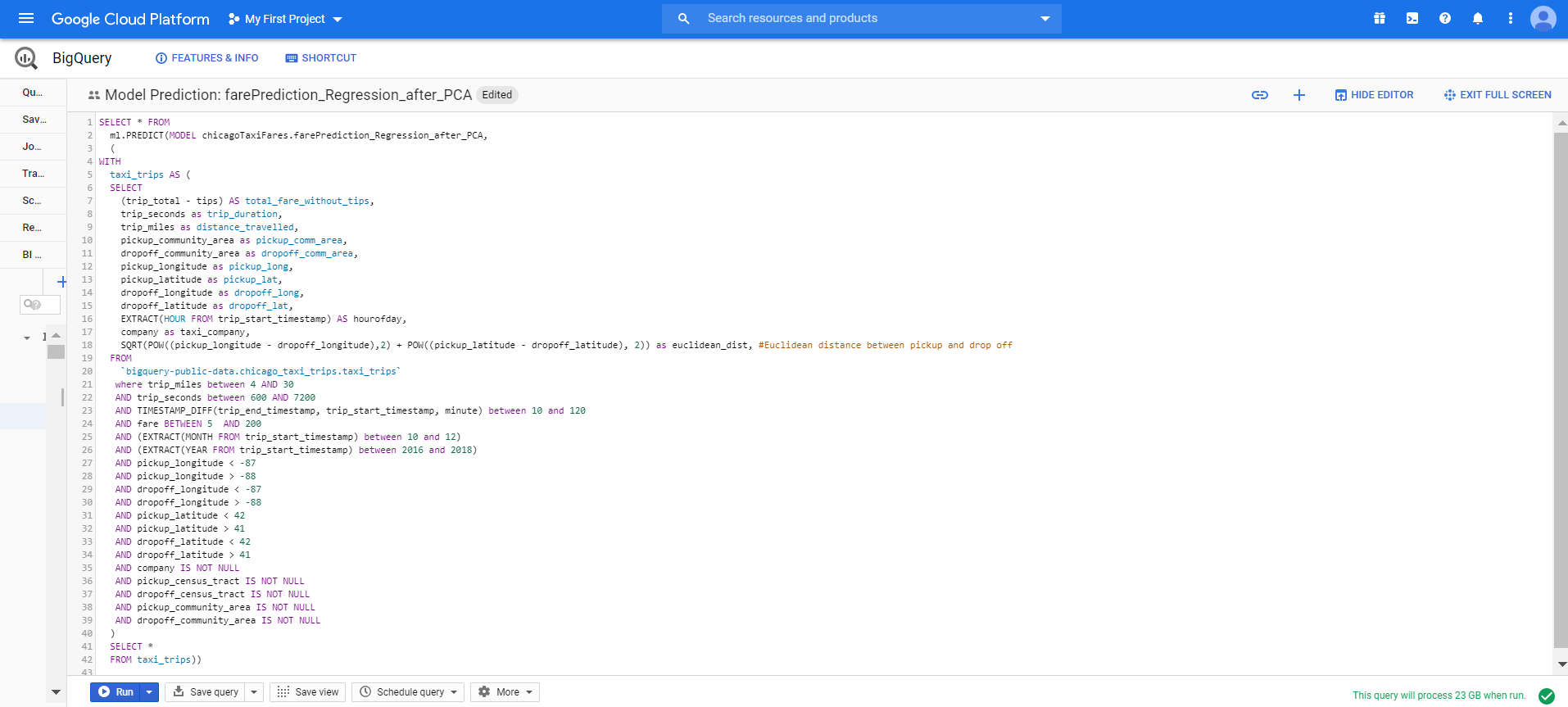
****

**Output Screenshot:**

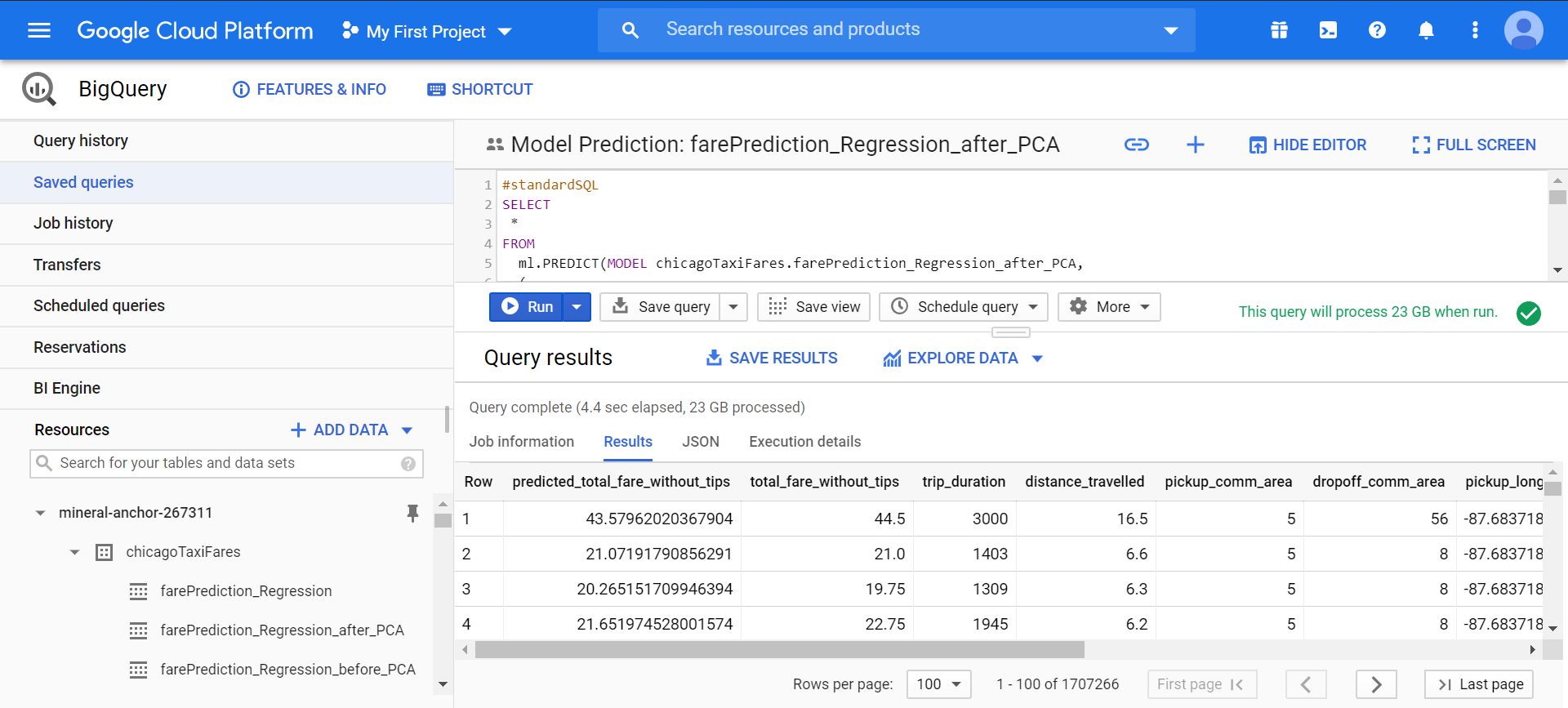
****

**6. Final Model - Prediction Query**

The prediction model uses training data from months 1 to 9 and the test data from months 10, 11, 12.

****

Below screenshot shows total\_fare\_without\_tips and the predicted\_total\_fare\_without\_tips using the final prediction model.

****