Since we have build a model on proter\_data.

Now we will answer few question related to this liner regression

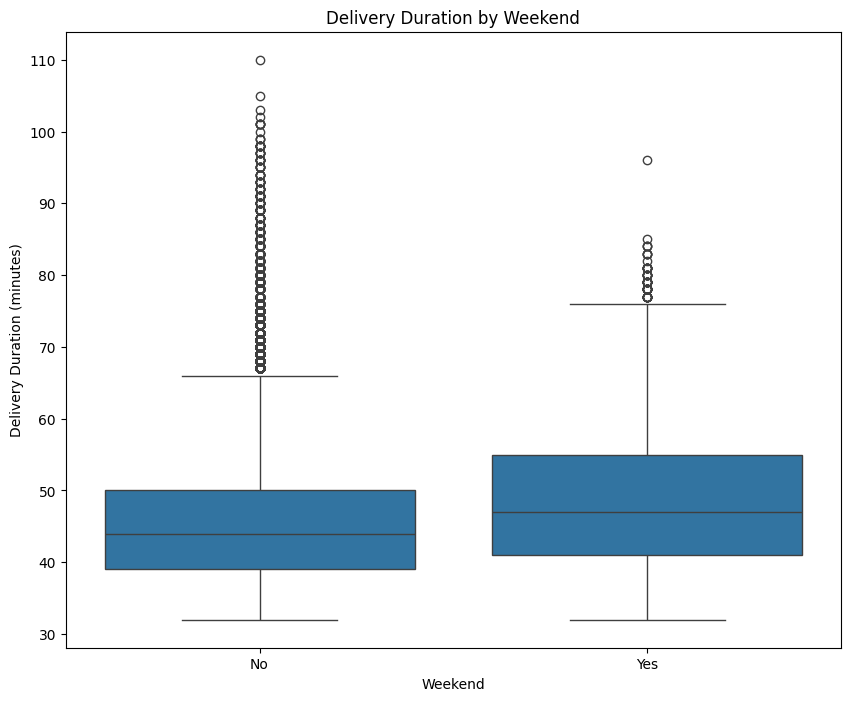
1. **Are there any categorical variables in the data? From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

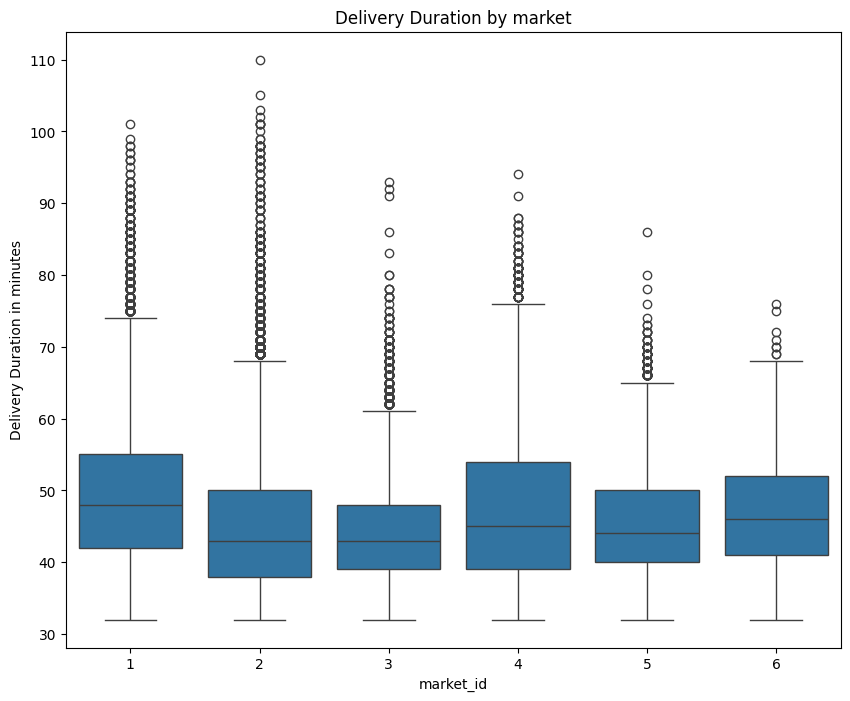
Ans: Yes, there are few categorical variables in this data. I found 3 in this data and they are A. market\_id, B. order\_protocol and C. weekend

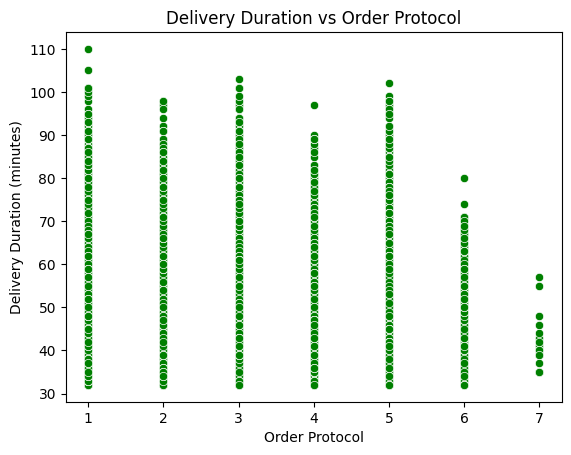
which we have created.

Yes, the categorical data are the impact in the dependent viable.

The graphs are on the python file. I am adding here also







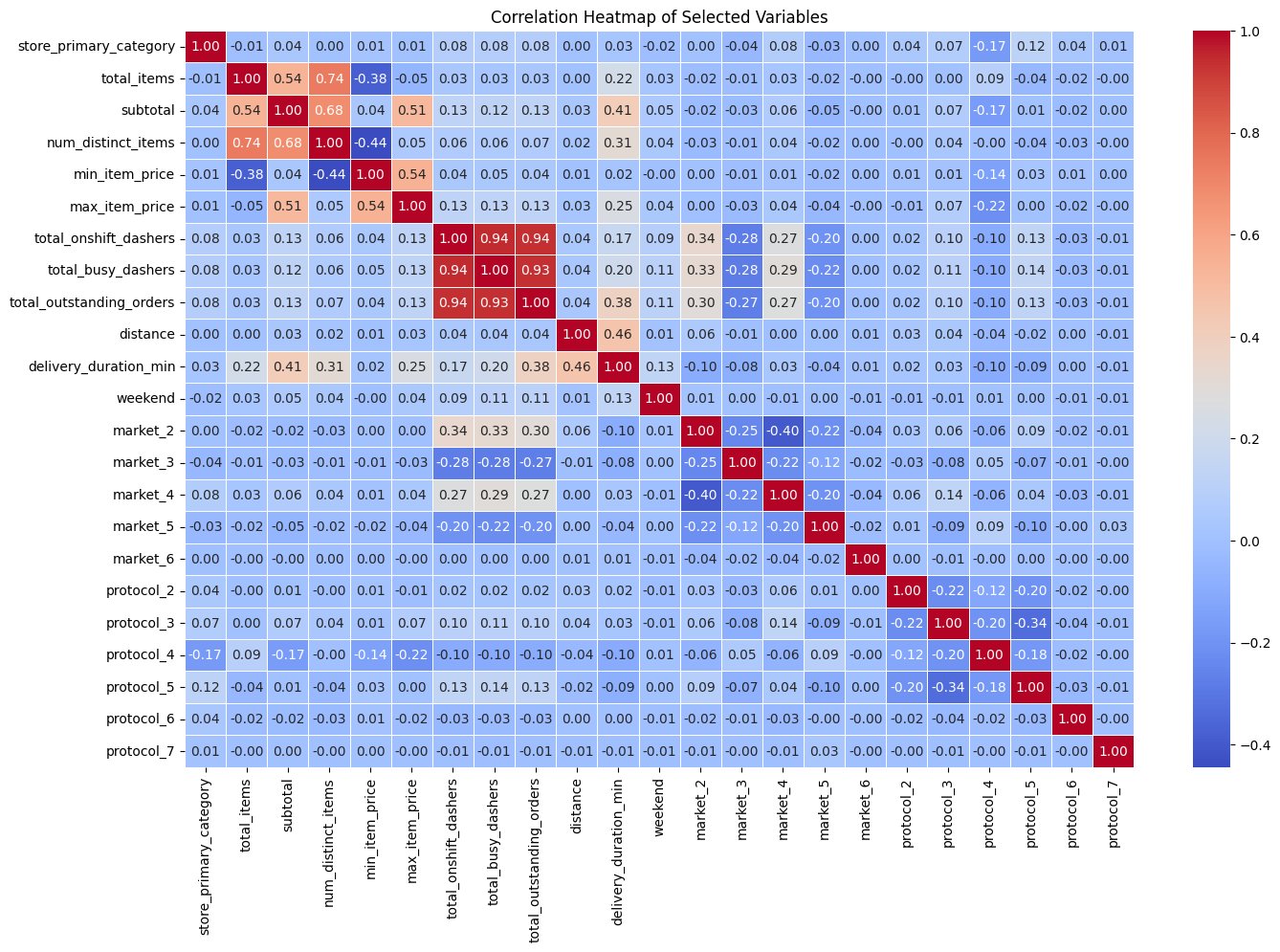
2. **What does `test\_size = 0.2` refer to during splitting the data into training and test sets?**

Ans: In above code are saying that we will set the test data containing 20% of total data

3.**Looking at the heatmap, which one has the highest correlation with the target variable?**

**Ans**: As per the heatmap we found that Distance is the highest co relation with the target variable

then we have another variables also

and the variables are 1. subtotal 2. total\_outstanding\_ordrs 3. num\_distnct  
Here is the heatmap as well   


4.**What was your approach to detect the outliers? How did you address them?**

**Ans:** we will take help of Boxplot graph

and we will do mimmax scale to take care of the outlies

**5. Based on the final model, which are the top 3 features significantly affecting the delivery time?**

**Ans:**

**6.Explain the linear regression algorithm in detail.   
Ans:**

**7.Explain the difference between simple linear regression and multiple linear regression**

**Ans: Simple Linear Regression- A model that uses one independent variable to predict a dependent variable.**

**Multiple Linear Regression- A model that uses two or more independent variables to predict the dependent variable.**

**8.What is the role of the cost function in linear regression, and how is it minimized?**

**Ans: In linear regression, the cost function measures how well the model’s predictions match the actual data.   
And the way to reduce it is Analytical Method (Normal Equation)**

**Linear regression has a closed-form solution using matrix algebra:  
9. Explain the difference between overfitting and underfitting.**

Ans: Overfitting and underfitting are two common problems that arise when building machine learning models. Overfitting occurs when a model learns the training data too well, including noise and random fluctuations, and performs poorly on new, unseen data. Underfitting, on the other hand, happens when the model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and new data.

**10. How do residual plots help in diagnosing a linear regression model?**

Ans : Residual plots are crucial for assessing the validity of a linear regression model by revealing patterns that suggest potential issues like non-linearity, heteroscedasticity, or the presence of outliers.