**DATABASE MANAGEMENT & BUSINESS INTELLIGENCE**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Course Project Report**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Devansh Shah**

**Rithika Mothukuri**

**Gautham Gonganda**

**Jennifer Indrupati**

**Data Management (MySQL)**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Description of the data:**

This dataset is for the company known as New-Wheels, a vehicle resale company, which has launched an app with an end-to-end service from listing the vehicle on the platform to shipping it to the customer's location. This app also captures the overall after-sales feedback given by the customer. The purpose of this is to answer complicated business questions and make them comprehensible for all the parties involved.

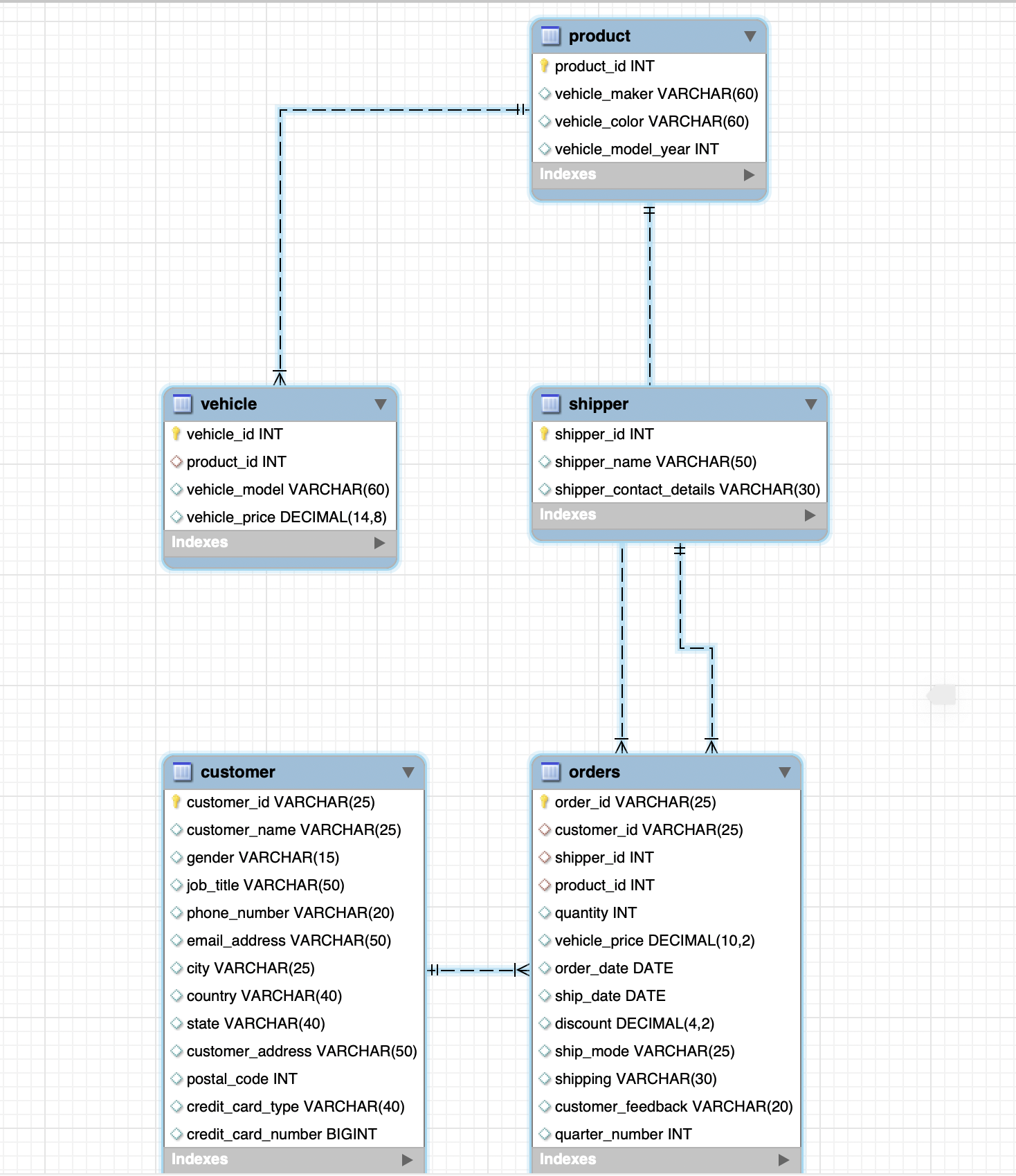
New-Wheels sales have been dipping steadily in the past year, and due to the critical customer feedback and ratings online, there has also been a drop in new customers every quarter, which is a problem to the business. We are going to look at key metrics to evaluate the state of the company.

**Data Description:**

We have 5 tables in the dataset.

* **Customers** table includes demographic data, and credit card information.
* **Orders** table includes all the information about orders placed like order data, shipping date, any discounts offered, and customer feedback on the order experience.
* **Product** table includes information about vehicle make, color and year
* **Vehicle** table includes information about vehicle model and price
* **Shipper** includes details about name of the shipping company, and their contact details

**ERD Diagram:**



**Queries and Results:**

**1). Customers are unhappy with shipping**

These customers rated their experience with the shipping company as either “Very Bad” or “Bad.” The difference between order date and shipping date is the largest among these customers as well, implying that long shipping times ruined their experience with the shipping company.

**2). Job titles with quantity of orders and vehicle makers is calculated**

This highlights a list of customers and their job titles and corresponding quantity of cars.

**3). Frequently used credit card type has been calculated**

Credit card types and the customers names who have used the cards have been calculated.

**4). Vehicles from the year 2006 have been retrieved along with their models:**

This implies that vehicles manufactured in the year 2006 have been extracted from the dataset, along with their respective models.

**5). The average price of vehicles for each model has been calculated:**

This indicates that the average price of vehicles has been computed for each distinct model.

**6). The most popular vehicle models ordered in the last year (2018) have been determined:**

This shows the vehicle models that were ordered the most frequently during the last year, which is 2023.

**7). States with highest average orders have been calculated:**

This provides a list states with highest average orders and the sales corresponding to the states.

**8). Total revenue generated from orders shipped by a specific shipper has been calculated:**

This calculates the total revenue generated from orders that were shipped by a particular shipper.

**9). Frequently used shipping mode, reviews and shipping companies are queried.**

This reveals most used shipping mode and whether customers are satisfied are by it or not.

**10). “Very Good” and “Good” ratings are only given to the fastest delivery options:**

This indicates that, customers are satisfied when the product is delivered very quickly.

**Conclusion:**

* We learned the New-Wheels has to reduce their time on shipment as that has been the cause of bad reviews. This suggests a need to review and possibly enhance the logistics and shipping strategies.
* Additionally, customers who provided the “Very Good” or “Good” feedback on the shipping experience chose either “first class” or “same day” deliveries. This implies that the shipping company needs to upgrade its business processes to enhance their delivery process since only the fastest delivery options are leaving customers satisfied.
* The occurrence of very bad feedback alongside orders with significant discounts might indicate issues such as customers' expectations not being met despite the discount, or quality problems with the discounted products.
* We also noticed that Maserati, GMC and Jaguar vehicles have the highest average order quantity. This suggests that New-Wheels should focus on marketing, stock allocation, and even production planning for Maserati.
* California, Minnesota, District of Columbia are the states with highest average orders as compared to the other states. This also is another direction where New-Wheels could head towards in-terms of marketing and stock allocation.
* Topdrive, the shipping company is the best out of all the shipping companies since their total revenue 183264.00
* We found that jcb credit card type has been used the most. New-Wheels can use this company’s information on their customers to plan targeted marketing campaigns.
* We found that customers in managerial positions ordered more cars as compared to customers in the middle and bottom level positions. This information could be useful in knowing which cars to target and where.
* We noticed that first class and same day deliveries are the most common but they also have bad customer reviews. This is something that New-Wheels should look into and improve on in order to increase customer satisfaction. We suggest that they change or drop shipping companies responsible for this: Katz, Rhybox, Tavu.

**Challenge and Lesson Learnt:**

* We came across a lot of issues when creating the tables itself. We learned that keeping the syntax constant is crucial. If the information is not in sync when it comes to creating tables, queries will not run; but from this issue we learned how to also fix them.
* When creating foreign and primary keys, it is better to create them while creating tables rather than after creating them by using the ALTER TABLE function. This simply created confusion and took a lot of time to understand and fix.
* We had to alter a lot of data since we had to create foreign keys. So we had to make sure that data was the same in all columns, for example, “customer\_id” in the tables customer and orders was not the same and therefore we were not able to create foreign keys between these tables. We learned how to quickly alter the table via the output tables and fixed the issue, but it took a lot of time as well.
* We also learned queries to check NULL VALUES in the data. This ended up being a huge advantage to learn since we had to alter a lot of our values and this function would tell us where information was not identical.
* Lastly, we learned a great deal about second-hand car companies and their workings. We also learned various price ranges for various car companies.

**Data Analysis using R**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Description of the dataset:**

We used the *Boston House Prices* for this part of the project ([link to Kaggle](https://www.kaggle.com/code/sagarnildass/predicting-boston-house-prices)). The intended purpose of this dataset is to evaluate the performance and predictive power of a model from data collected from homes in suburbs of Boston, Massachusetts. The model, if turns out to be a good fit, can be then used to make predictions about a house, particularly, its monetary value. This model is also quite valuable for someone like a real estate agent.

Since the dataset only has 4 variables, one being the target variable, we decided to not cut down on any part of the dataset for analysis. We decided to use the entire dataset here, apart from treating outliers, which will be explained below. Additionally, the entire dataset is numeric, there are no categorical variables.

Following is the description of the variables in the dataset:

* **RM** is the average number of rooms among homes in the neighborhood
* **LSTAT** is the percentage of homeowners in the neighborhood considered lower class / working poor
* **PTRATIO** is the ratio of students to teachers in primary and secondary schools in the neighborhood.
* **MEDV** is the target variable that we are trying to predict; it is simply the price of the houses.

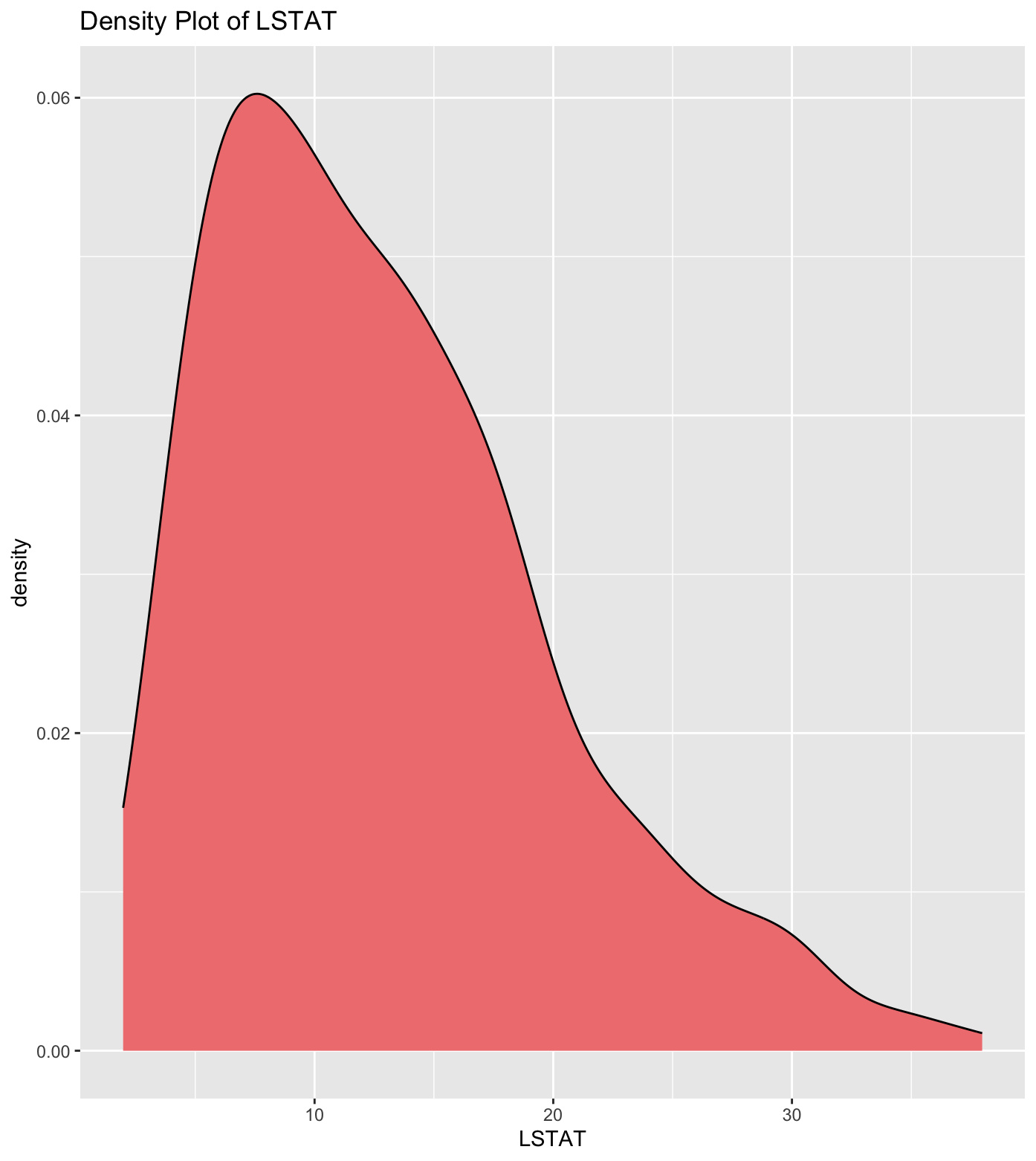
**Descriptive Statistics:**

* Using the summary() command, we got an idea of how the dataset works. The average price on houses was around $440000; ranging from $105000 to $1024800.
* The PTRATIO ranges from 12.60 to 22
* RM (average number of rooms) ranges from 3 to 8, the average being 6.
* Average homeowners considered to be in the lower class are around 12% (LSTAT variable).

**Graphing the Variables:**

We decided to graph all the variables using the ggplot package and graphed them into histograms as well as density plots.

Findings:

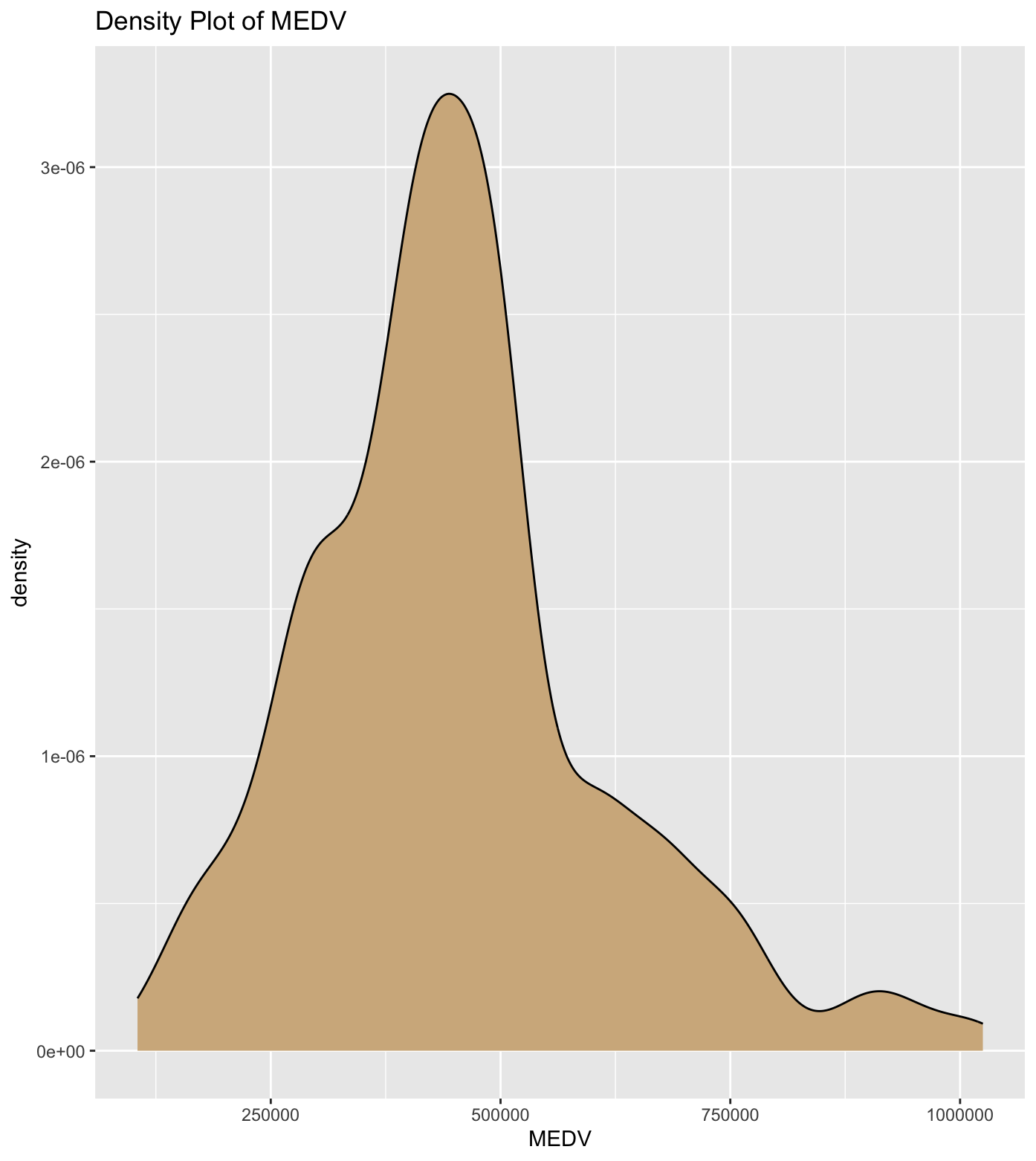


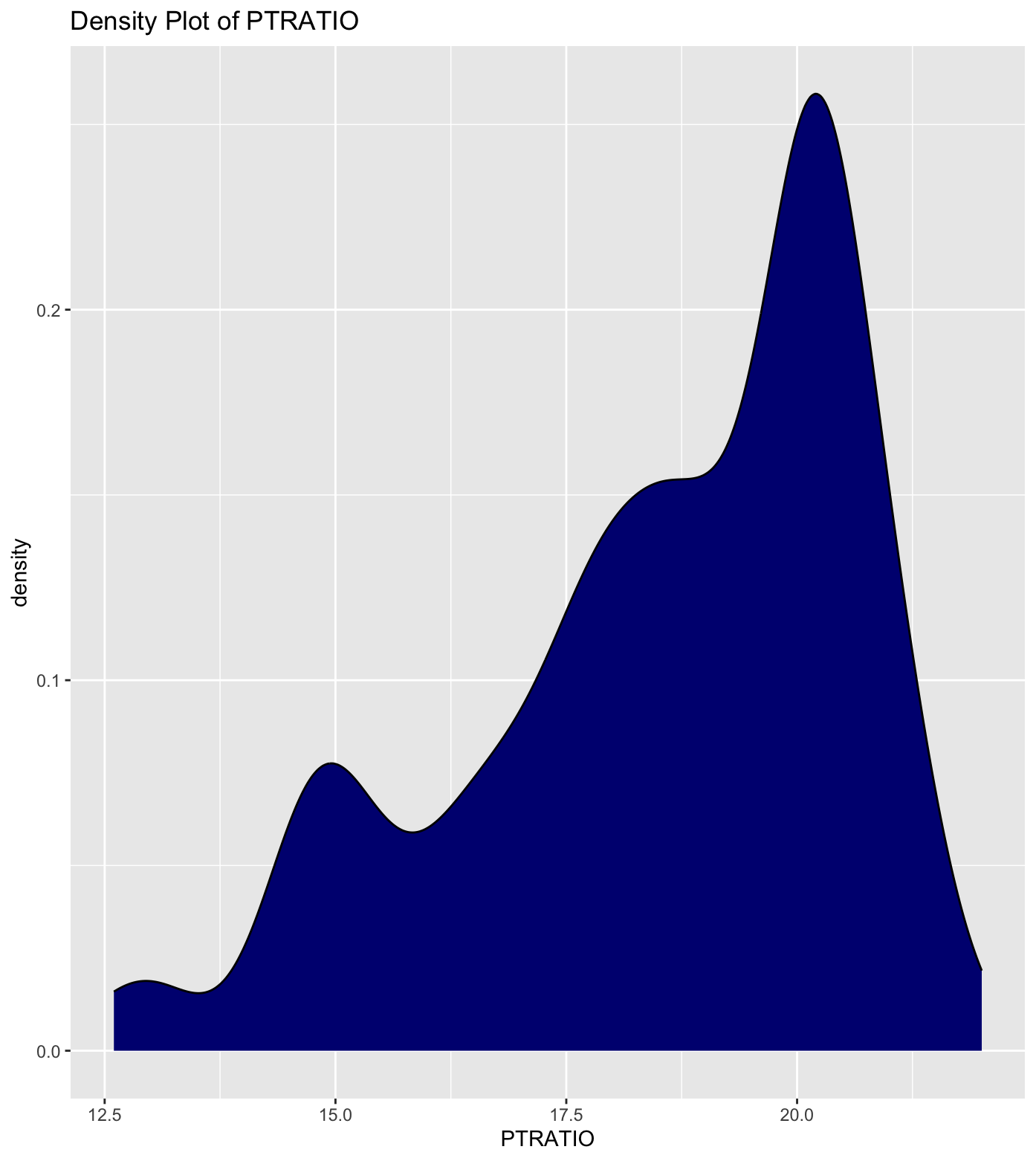
Min. 1st Qu. Median Mean 3rd Qu. Max.

1.98 7.37 11.69 12.94 17.12 37.97

According to the graph and the quantile distribution obtained via the summary() command above, the LSTAT variable was right skewed, and had a large variation in the distribution of its units. This meant that there were outliers in it which we decided to treat later in the analysis.

The rest of the variables seemed relatively normally distributed with not a significant difference in their distribution.



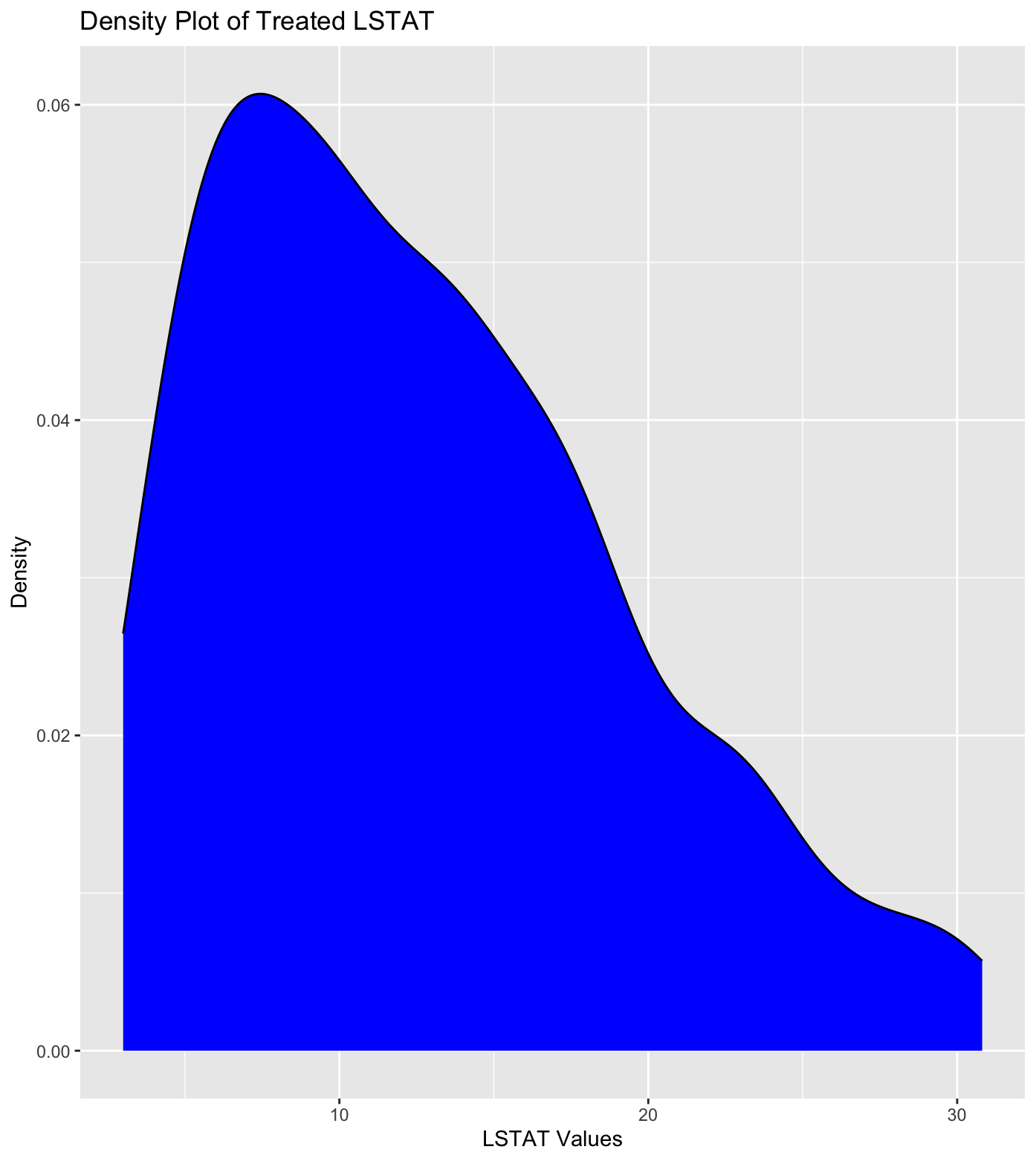


**Treating the outliers:**

One of the assumptions in regression is that any data is usually normally distributed, but that does not usually happen in real-life datasets. As mentioned earlier, the LSTAT variable in the dataset was right skewed, therefore, we decided to treat it to remove the outliers. Treating outliers improves the regression result and therefore, improves the predictive power of the model. To treat the variable, we decided to omit any numbers outside the range of 3 and 30. Post-treatment, the graph showcased a better distribution and the variation in distribution also decreased, as shown below.

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.00 7.37 11.69 2.77 17.12 30.81

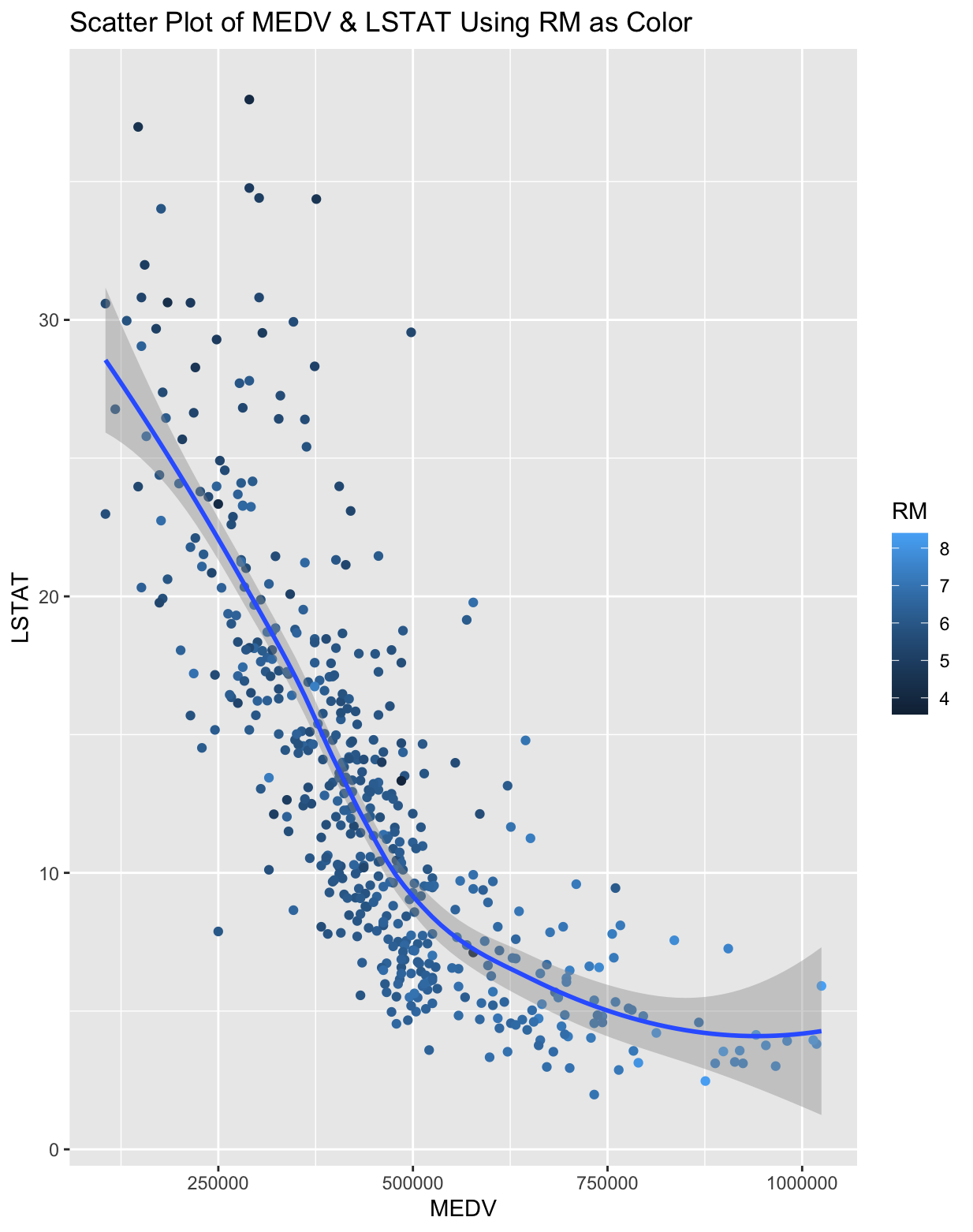
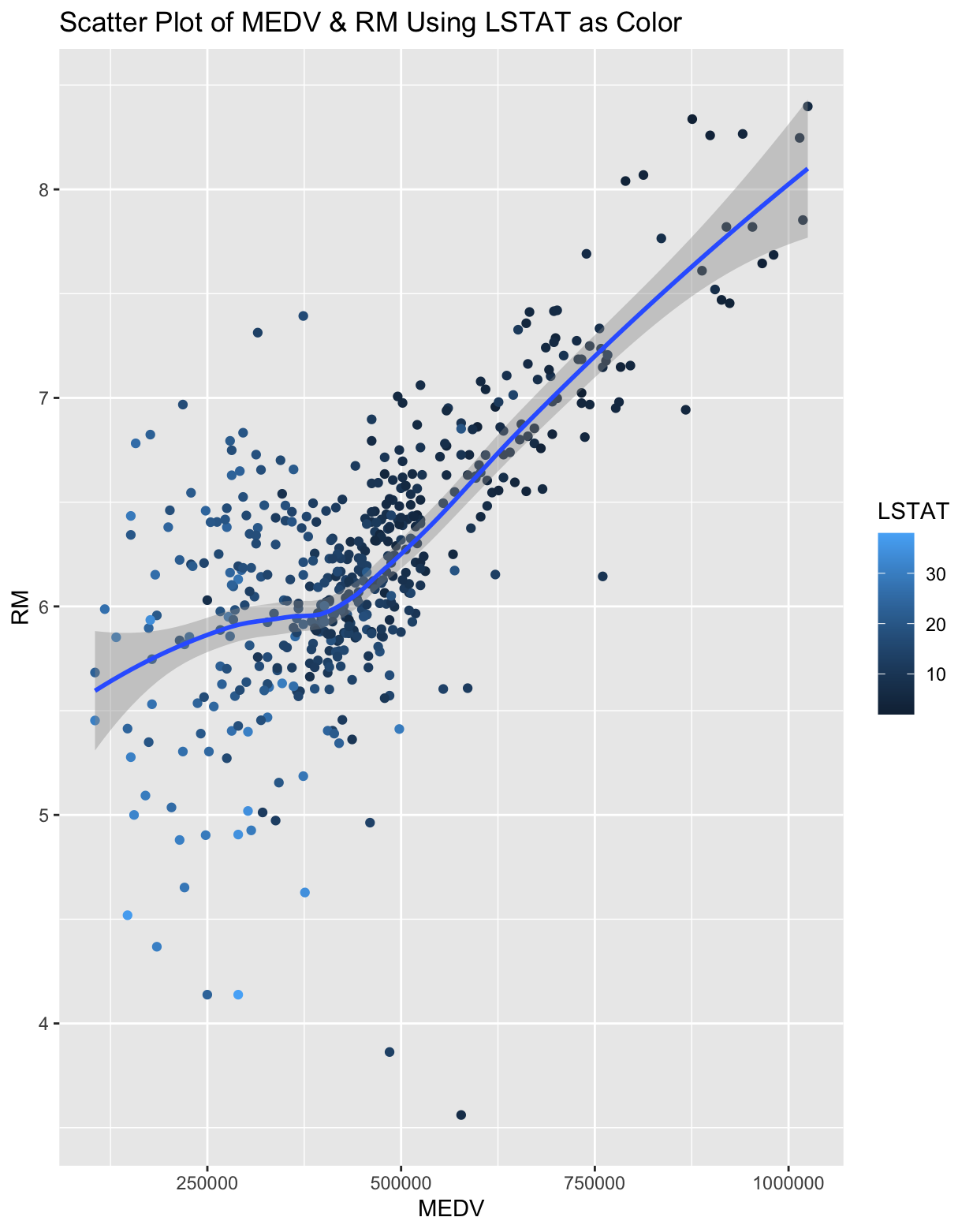


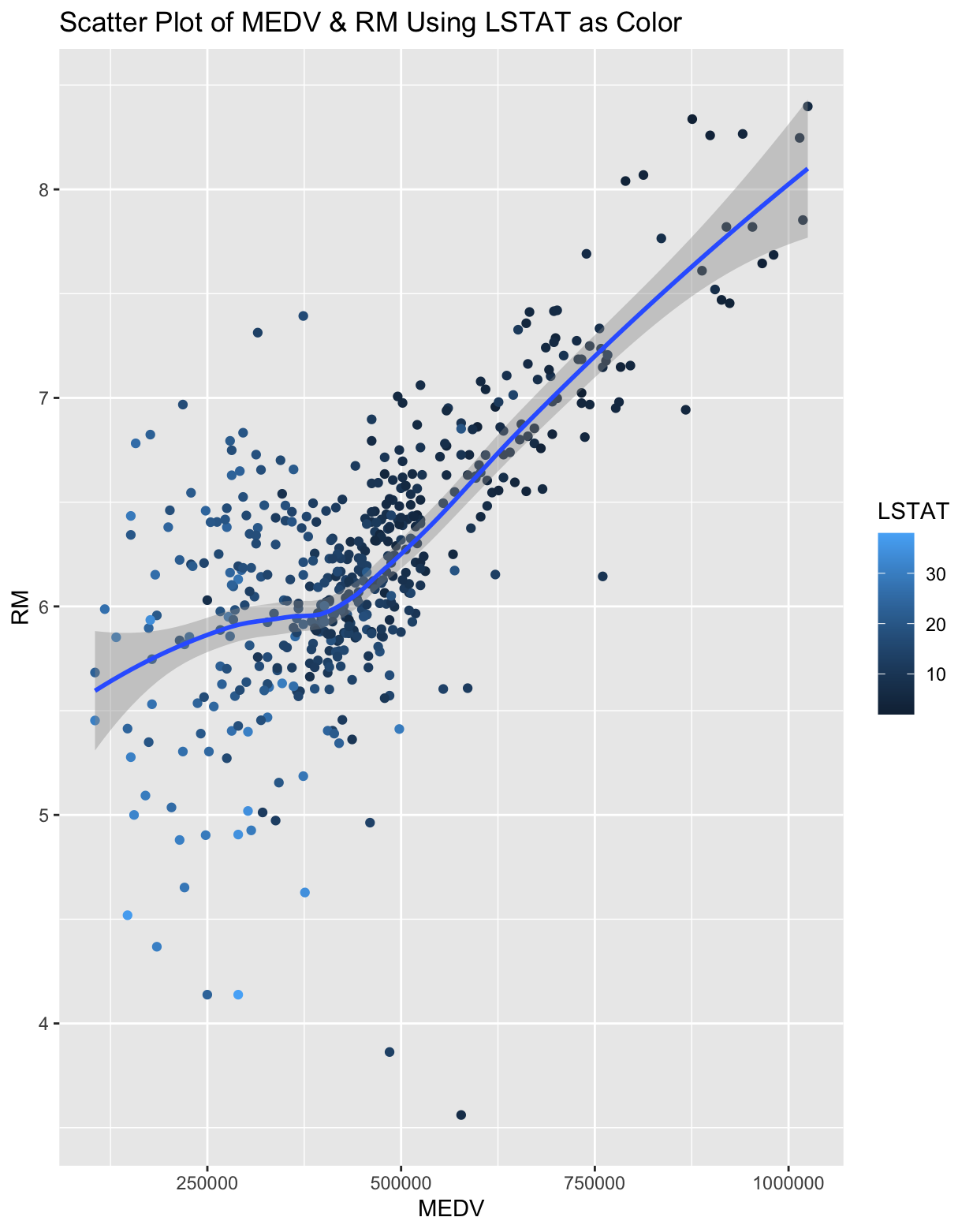
Lastly, even though the PTRATIO graph was left skewed, we chose not to treat because the variation in distribution did not have a large difference:

Min. 1st Qu. Median Mean 3rd Qu. Max.

12.60 17.40 19.10 18.52 20.20 22.00

**Correlation:**

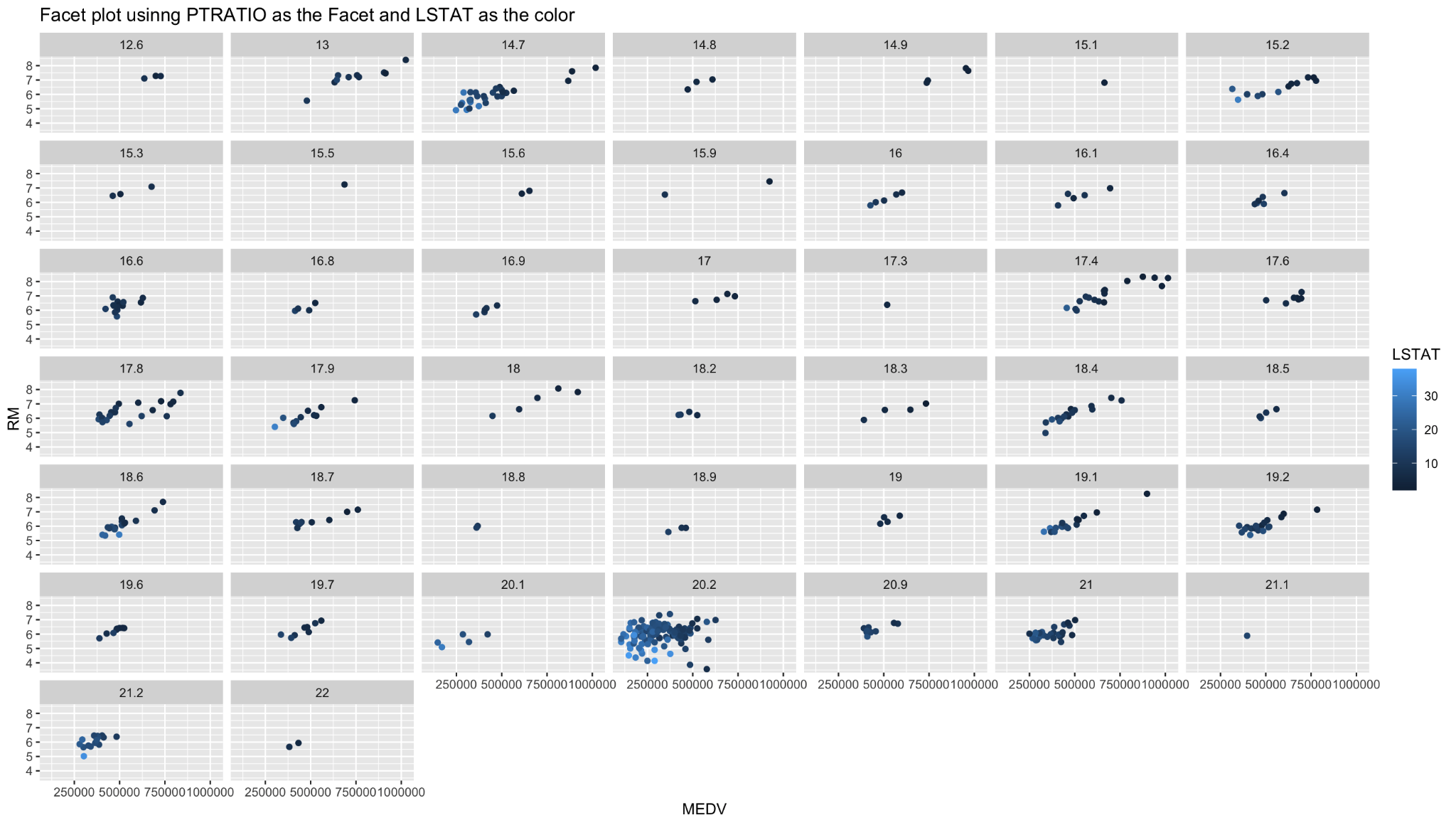
****

****

* We found that MEDV and LSTAT had the highest negative correlation (-77%). This means that the more the value of LSTAT, the less will be the value of MEDV. It is pretty evident that with the increase in the lower class homeowners, then it is likely that very expensive real estate owners will not build their housing complexes in that region as most of the people will not be able to afford it.
* RM and MEDV had the highest positive correlation (69%). This means the more the value of RM, the more will be the value of MEDV. It is pretty evident that with an increase in the number of rooms, the price of the house will increase.
* Lastly, we observe that PTRATIO and MEDV are negatively correlated (-51%). This implies that the more the value of PTRATIO, the less will be the value of MEDV. It is pretty evident that with an increase in the students to teachers ratio, teachers will not be able to attend to students individually everytime and hence this may affect the education of students. So regions with a low PTRATIO will have higher prices for houses.
* Correlation between the other variables was not as significant.

**Faceting:**

Further, we plotted the variables using scatter plots to understand them better. We used the facet\_grid function and used PTRATIO as the facet.



* Above graph is a proof to show the behavior of the target variable with respect to other features.
* As an example, we see that the price increases as the lower income group percentage decreases.
* Moreover, as the student to teacher ratio increases (PTRATIO), price also decreases, as we also saw previously.
* There is also a large concentration of data points in the ratio section where for every 20 students or so, there is one teacher.
* Lastly, we also came to an understanding that since there are not a lot of data points here, it also showcases that having more data points would enable better understanding.

**Regression Analysis:**

**Linear Regression**

**Hypothesis:**

Since the RM variable has the highest positive correlation with MEDV we assumed that it would be statistically significant and have an impact on pricing.

Note:

We also kept in mind that since the dataset is small, the linear regression using single variables might not provide good results.

Multiple R square: 48.61%

Adjusted R square: 48.5%

P-value: < 2.2e-16

As we can see, the R square is not an extremely good number and this could be caused by the size of the dataset. And it could simply mean that all the variables combined would give a much better result.

We also conducted a regression analysis between MEDV and LSTAT, this too gave a lower R square.

**Multiple Regression**

**Hypothesis:**

All the three predictor variables will have a significant impact on the pricing/target variable (MEDV).

Multiple R square: 73.75%

Adjusted R square: 73.59%

P-value: < 2.2e-16

* This is a much bigger improvement compared to the linear regression.
* This means about 72% of variation in target variable (MEDV) can be predicted by predictor variables, which is quite good.
* The p-value is also much lower than 0.05, implying the statistical significance of the model.

Regression equation:

MEDV = 423852.94 + (85726.9 \* RM) (-12162.8 \* LSTAT) (-18857 \*PTRATIO)

**Conclusion:**

* If the LSTAT and the PTRATIO were fixed, on average RM will be costing 85726.9 more than others.
* The coefficient of about -12162.8 for LSTAT tells us that for a given RM and PTRATIO, the predicted MEDV decreases by about -12162.8 for every 1.0 unit increase in LSTAT.
* The coefficient of about -18857 for PTRATIO tells us that for a given RM and LSTAT, the predicted MEDV decreases by about -18857 for every 1.0 unit increase in PTRATIO.

**Challenges:**

* While graphing the variables, we couldn’t use the face\_grid function while graphing these variables because the graph was too huge and complex to look at and understand.
* Since the dataset did not have any categorical variables, we decided to use PTRATIO in the face\_wrap function and it gave us a comprehensive perspective. Using other variables as facets under this function distorted the results.
* In the same vein, we thought that having more variables would give us more insights into what affects house pricing, like the location - different suburbs acting as categorical variables, accessibility in the neighborhoods in terms of grocery stores, malls, schools, etc., driving distances, for example would have been helpful.
* While writing a SQL query in R, we realized that using double quotes in the query for the WHERE clause confuses the system since the entire query has to be in double quotes. To tackle this problem, we used single quotes in the WHERE clause which worked.