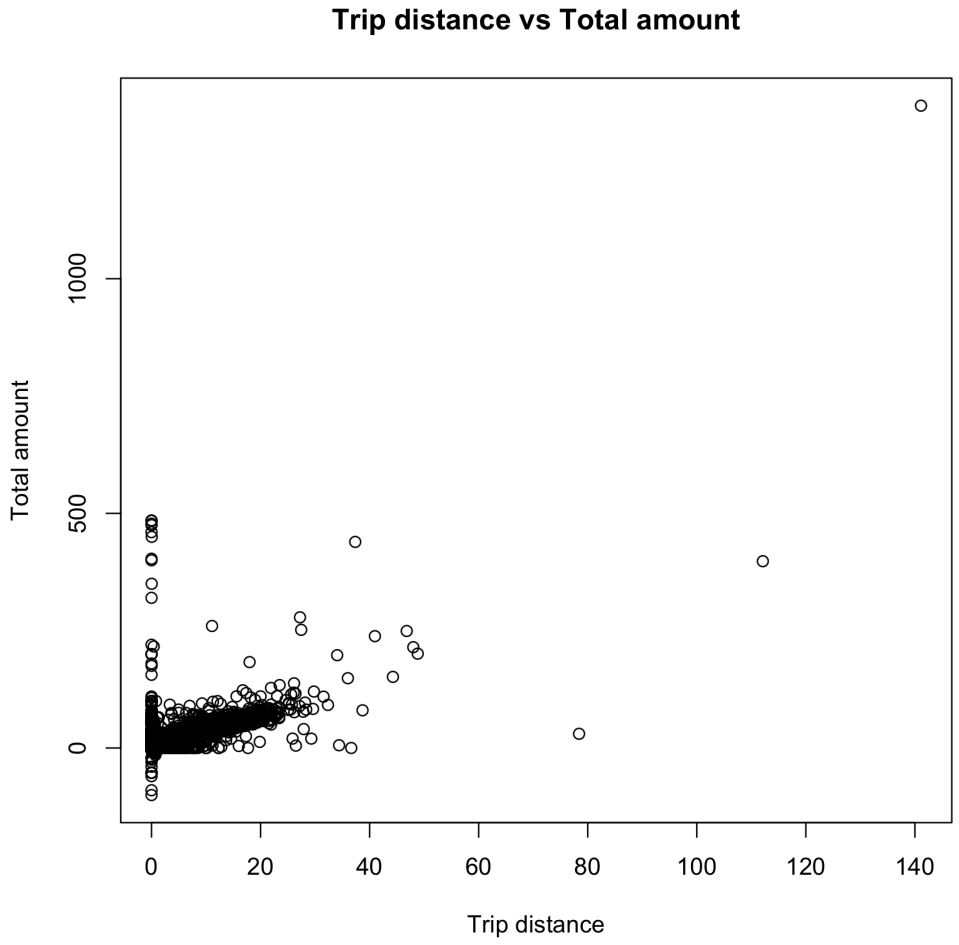
Name: Vikram Sahai Saxena
Net ID: vs799
RUID: 219004709

Module 4 Exercise: Scatter Plots

i. **The entire R code used when creating the scatter plot in (1)**

```
greencab<-read.csv("greencab.csv")
plot(greencab$Trip_distance, greencab$Total_amount, main="Trip distance vs Total amount", xlab="Trip distance", ylab="Total amount")
```

ii. **Screenshot of the scatter plot created in (1)**



iii. **Reasoning for selecting variables to be depicted in axis**

Trip distance is considered an independent variable because it's something that is determined before the total amount is calculated. So, it is selected to be depicted on x-axis.

The total amount is considered a dependent variable because it can be influenced by the trip distance. So, it is selected to be depicted on y-axis.

iv. The message being sent out by this plot

The plot aims to visualize any potential correlation between the trip distance and the total amount charged for the trip. For example, a positive correlation where points trend upward as you move right would suggest that longer trips tend to cost more, which is an expected outcome in most fare structures.

v. Weaknesses in this plot

- Overplotting: If the dataset is large, points may overlap, making it hard to discern the density of points in areas of the plot.
- Outliers: Extreme values can distort the scale of the plot, making it difficult to observe the general trend.
- Context: Interpreting the relationship between trip distance and total amount may be difficult without additional context or details, like the units of measurement or the scale used.
- Limited precision: Scatter plots primarily reveal visual trends and do not provide exact conclusions, which restricts the capability to extract in-depth insights directly from the plot.

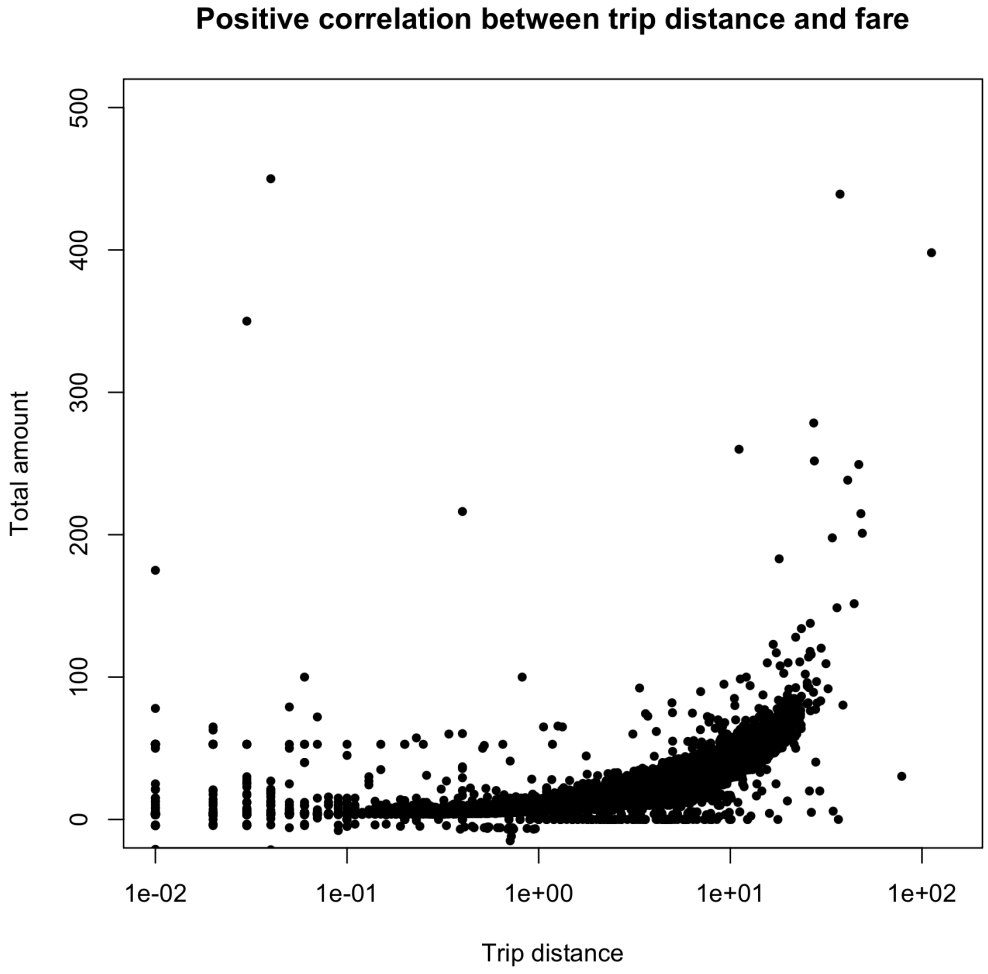
vi. List of actions to minimize the identified weaknesses

- Highlight Different Data Segments: Use distinct colors or shapes to differentiate between various data segments or categories, revealing hidden patterns or trends that might be lost in the overall data.
- Modify Scale: For data that is unevenly distributed or covers a broad range, applying a logarithmic scale to the axes can lead to a truer depiction.
- Augment with Supplementary Data: Improve visualization by incorporating additional elements like trend lines or mean values, thus providing more context and aiding in the understanding of the key trends or connections.
- Consider Different Plot Styles: Consider other types of visualizations like line graphs or histograms to more effectively convey the relationship between variables.
- Implement Error Bars: Error bars should be added to reflect the variability or uncertainty in the data, offering a better understanding of the data's consistency and the correlation's robustness.

vii. The entire R code used when creating the scatter plot in (6)

```
plot(greencab$Trip_distance, greencab$Total_amount, log="x", ylim=c(0,500),
main="Positive correlation between trip distance and fare", xlab="Trip distance",
ylab="Total amount", pch=20)
```

viii. Screenshot of the scatter plot created in (6)



ix. The entire R code used when creating the scatter plot in (7)

```
plot(greencab$Trip_distance, greencab$Total_amount, log="x", ylim=c(0,100),
main="High fares charged for certain short distance trips", xlab="Trip distance",
ylab="Total amount", pch=20)
```

x. Screenshot of the scatter plot created in (7)