**Name : Ganesh Arunagiri Rajan**

**Student ID : 13662388**

**Assessment task 1 - part A: regression modelling**

**Introduction** **:**

This report deals with the analysis performed and insights gained from the transactions dataset while trying to predict the monthly sales for December 2016 for every Industry across all locations. The CRISP-DM methodology was followed at all stages of analysis for predicting the monthly sales, which includes :

* Business Understanding
* Data Understanding
* Data Preparation
* Modelling
* Evaluation
* Advanced Model Fitting

**Business Understanding :**

The client who is a Sales Manager wants to predict if their organisation would have an increase in December Christmas Sale revenue from existing customers at all locations.

**Data Understanding :**

The provided dataset contains 94248 records of sales transactions made by 4464 unique customers across all the 89 industry location combinations from January 2013 to November 2016. Five different parameters for each and every transaction has been recorded which includes date, customer\_id, industry, location and monthly\_amount.

Out of the 89 Industry-location combinations, only 75 combinations had transactions on all the 47 months during the time period from January 2013 to November 2016. However, 5 combinations had transactions on lesser than 20 months.

Upon aggregating the data to location-monthly level, it was observed that the mean monthly amount was following an upward trend for all locations except location 1 and 8 where a downward trend was observed(Fig 1).

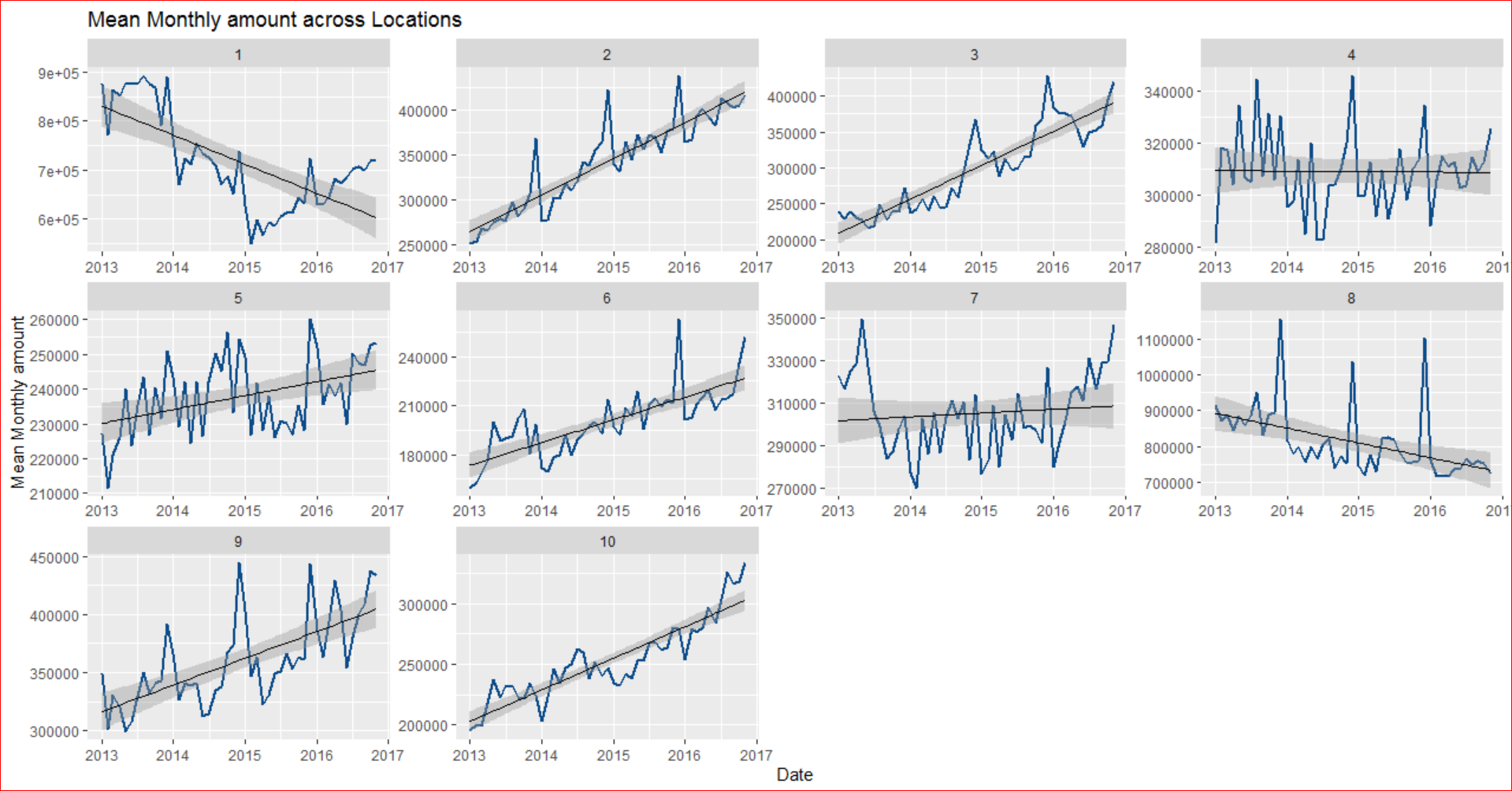


Fig 1: Mean monthly amount across locations.

At the same time when aggregated at a Industry-monthly level, an increasing trend of mean monthly amount was observed for all industries except industry 5(Fig 2).

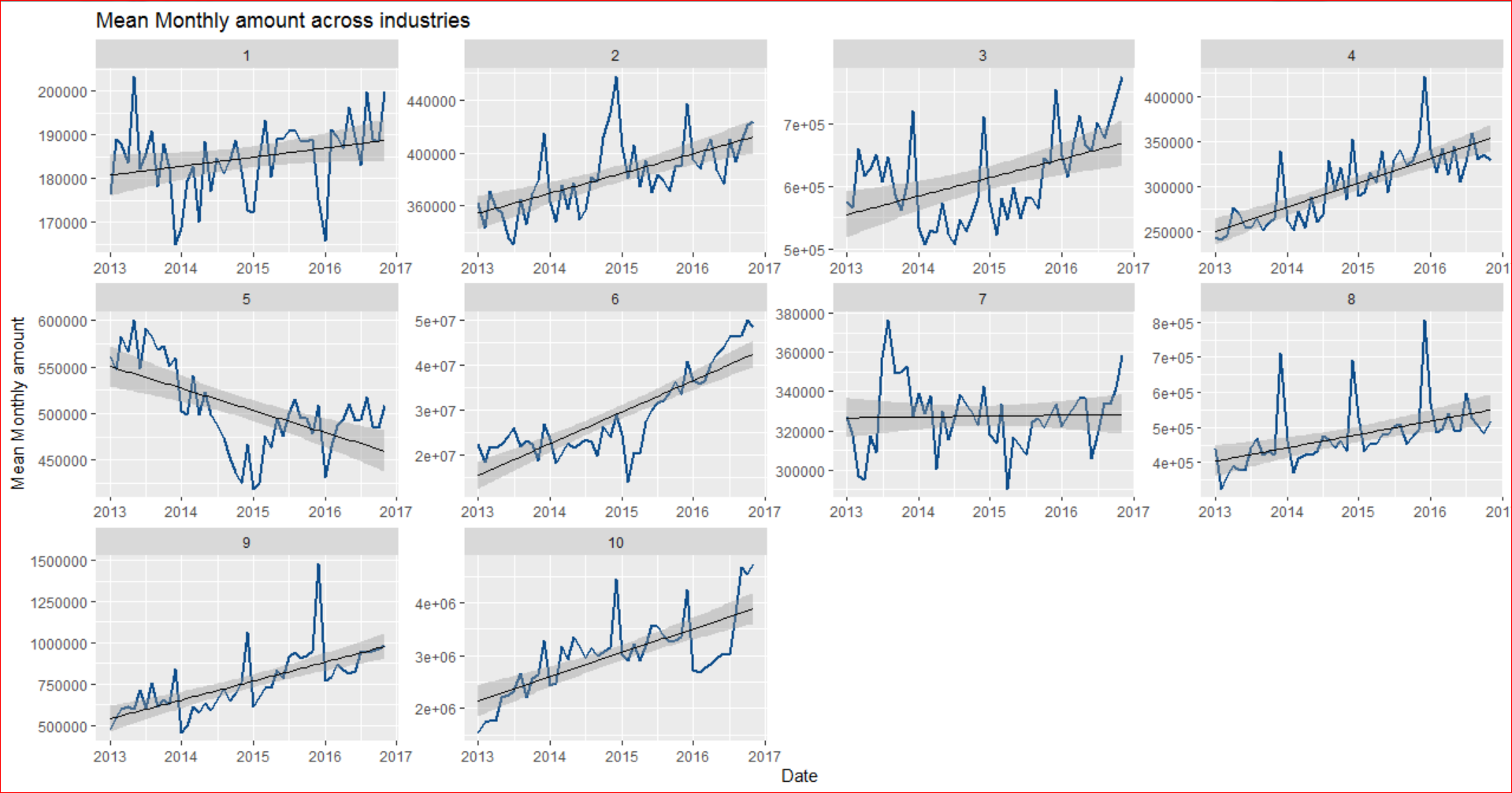


Fig 2: Mean monthly amount across industries

Moreover in the Industry-monthly aggregation, an increasing trend of customers was observed across all locations except Industry 4, 6 and 10(Fig 3).

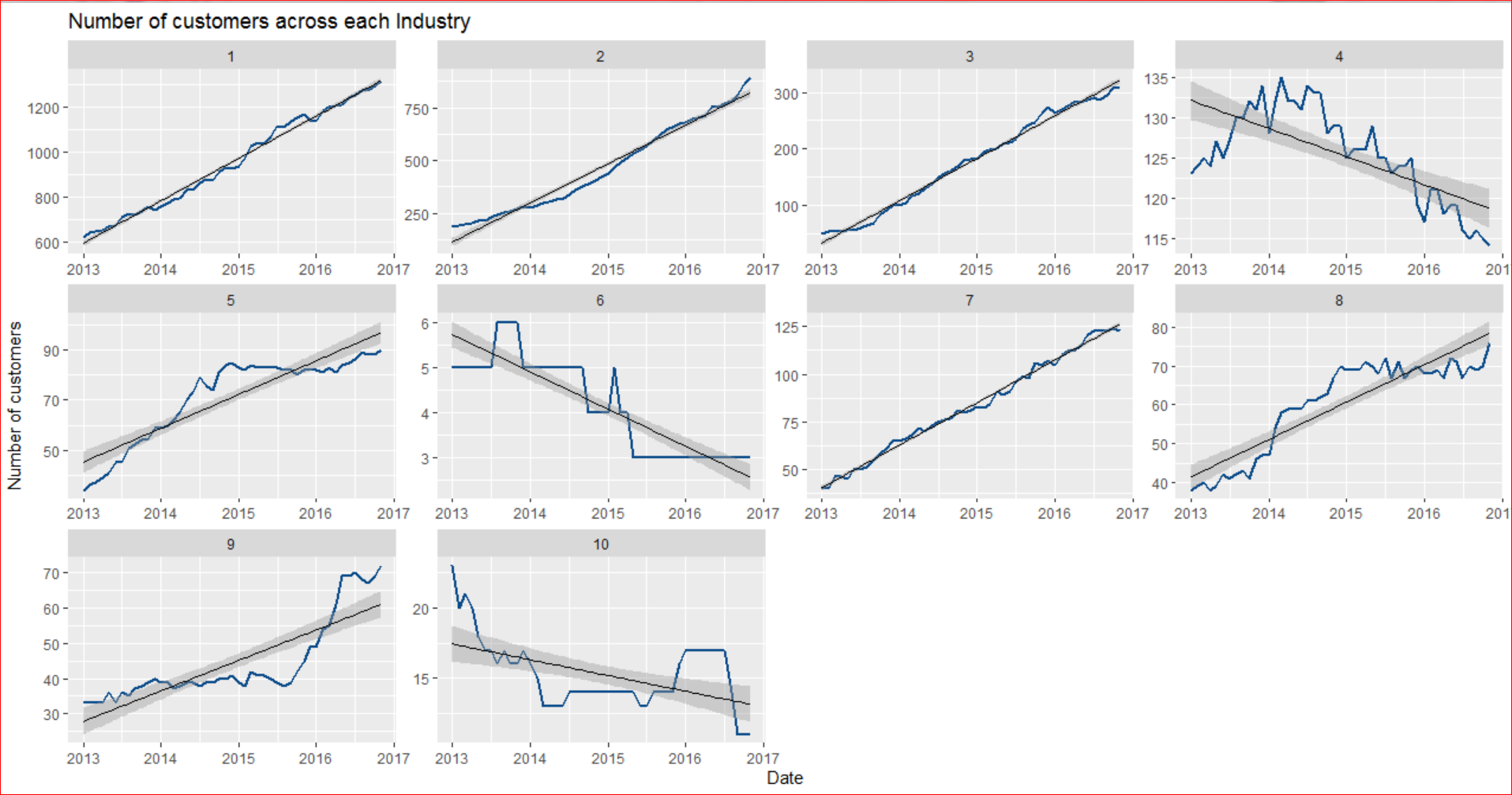


Fig 3: Number of customers across each Industry

However for the location-monthly aggregation, an increasing trend was observed across all locations(Fig 4).

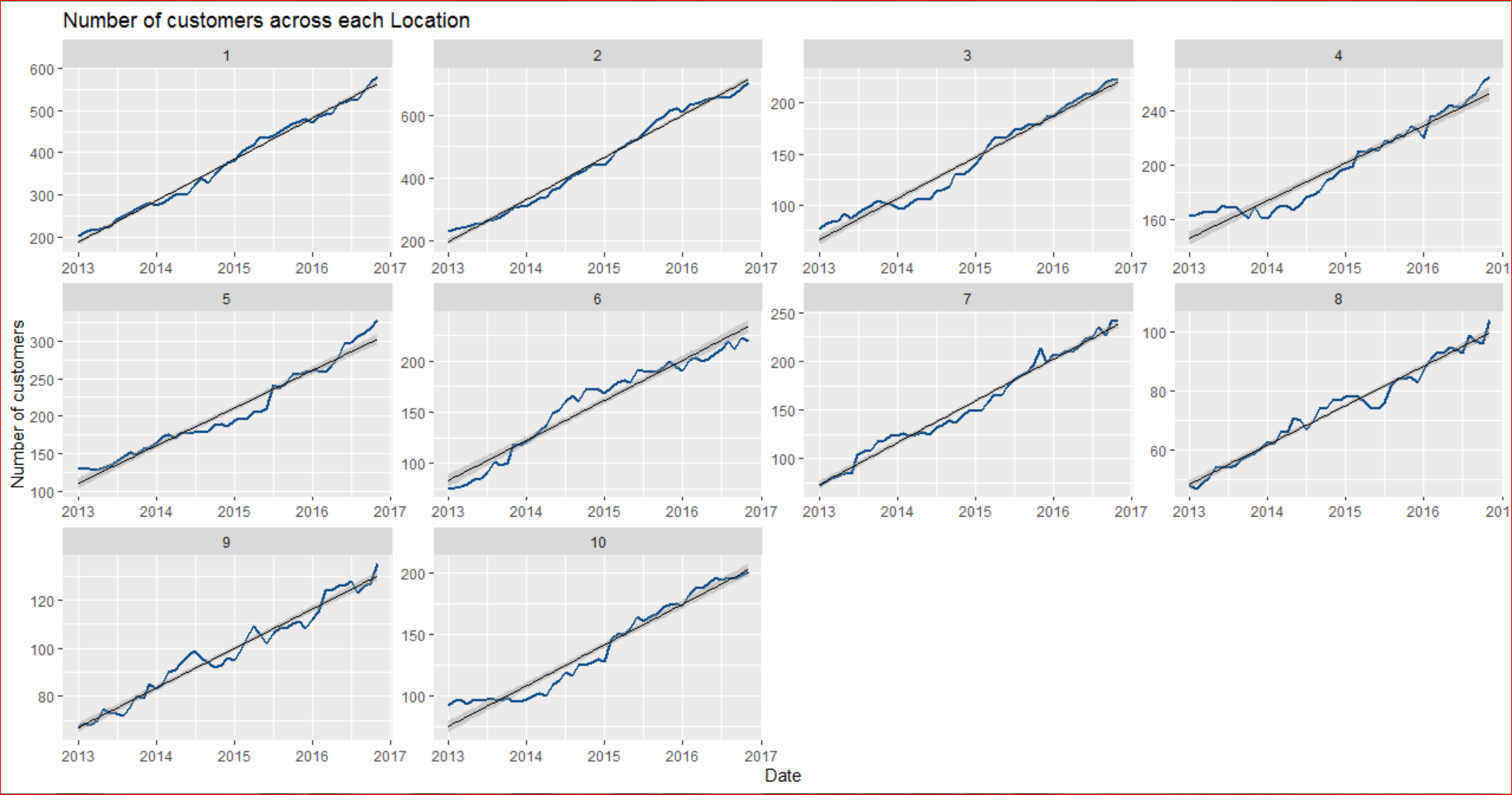


Fig 4: Number of customers across each location

From the above observations it is interesting to note the following :

* Despite an increase in the number of customers the average monthly sale amount of locations 1 and 8 is decreasing(Fig 1 & 4).
* The average monthly amount of Industry 5 is decreasing despite an increase in customers(Fig 2 & 3).
* Although Industries 4,6 and 10 face a drop in customers their mean monthly sales amount is increasing(Fig 2 & 3).

Also, Industry 2 and location 1 have grossed the highest total amount in their respective categories for the provided time frame(Fig 5 & 6).

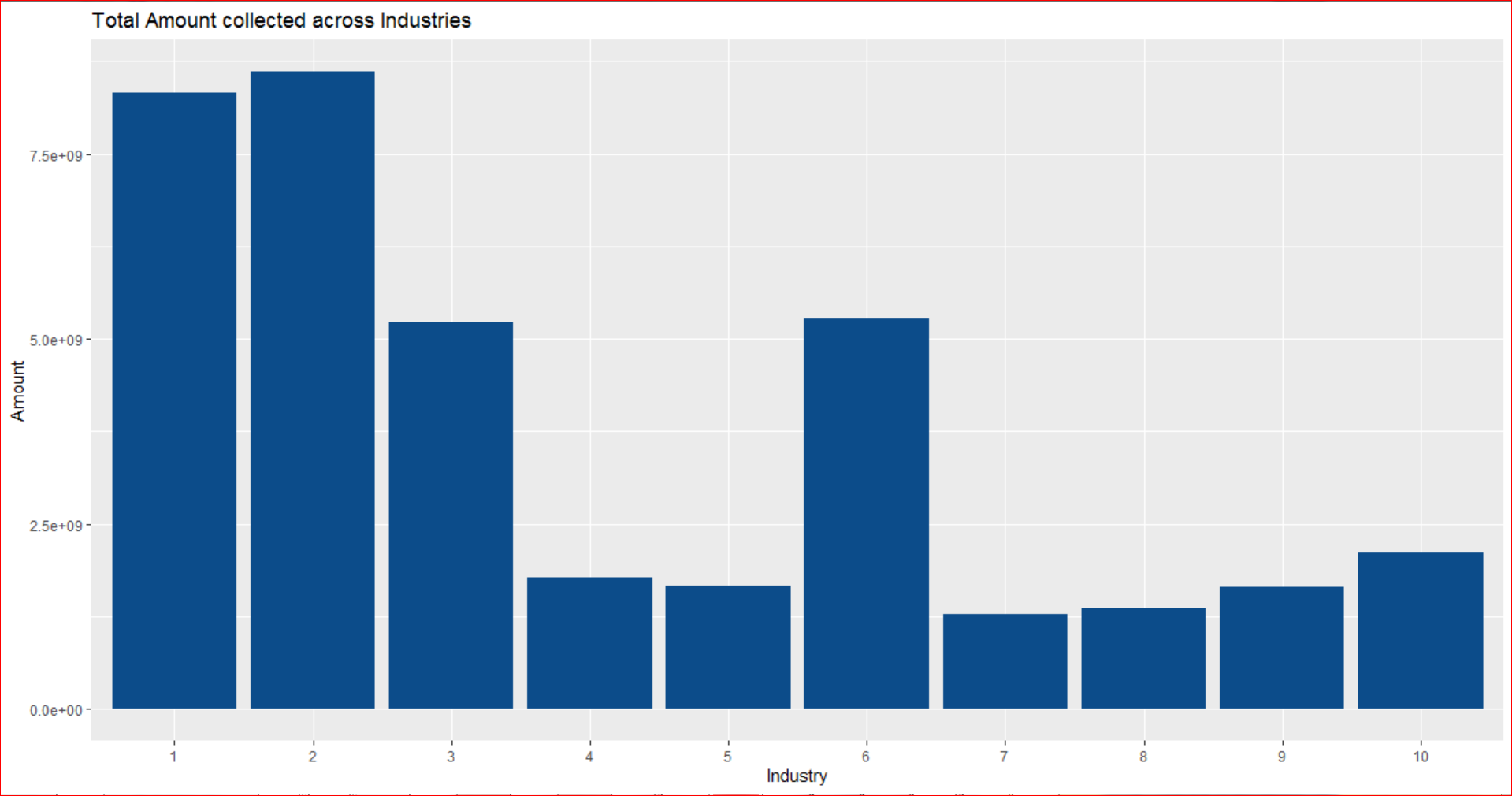


Fig 5 : Total amount collected across industries

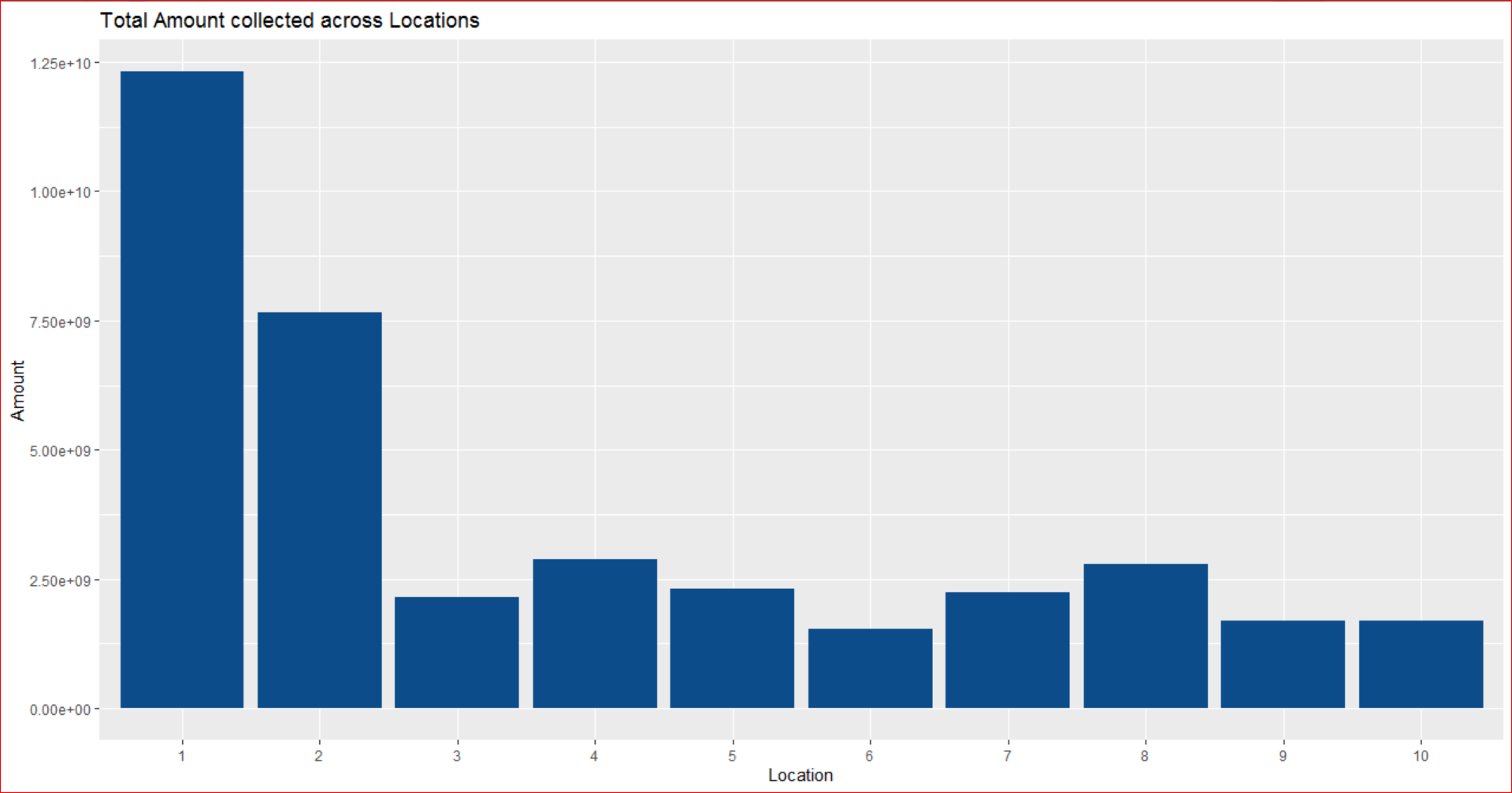


Fig 6 : Total amount collected across locations

Also the highest total monthly collections were observed in Month 11 - November (Fig 7)

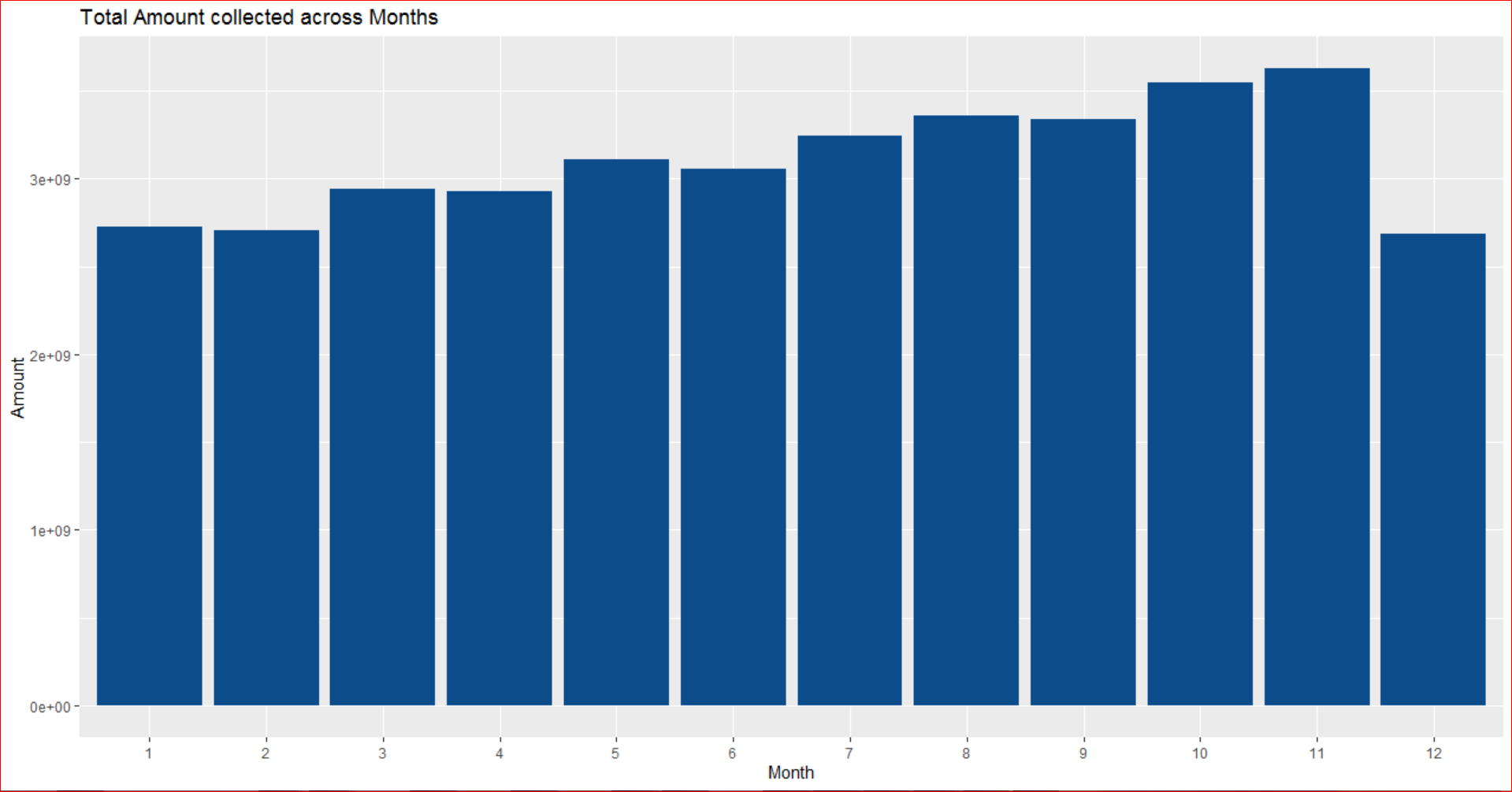


Fig 7 : Total monthly amount collected across months

Industry 6 - location 1(combination) has grossed the highest total amount from 2013 to November 2016(Fig 8).

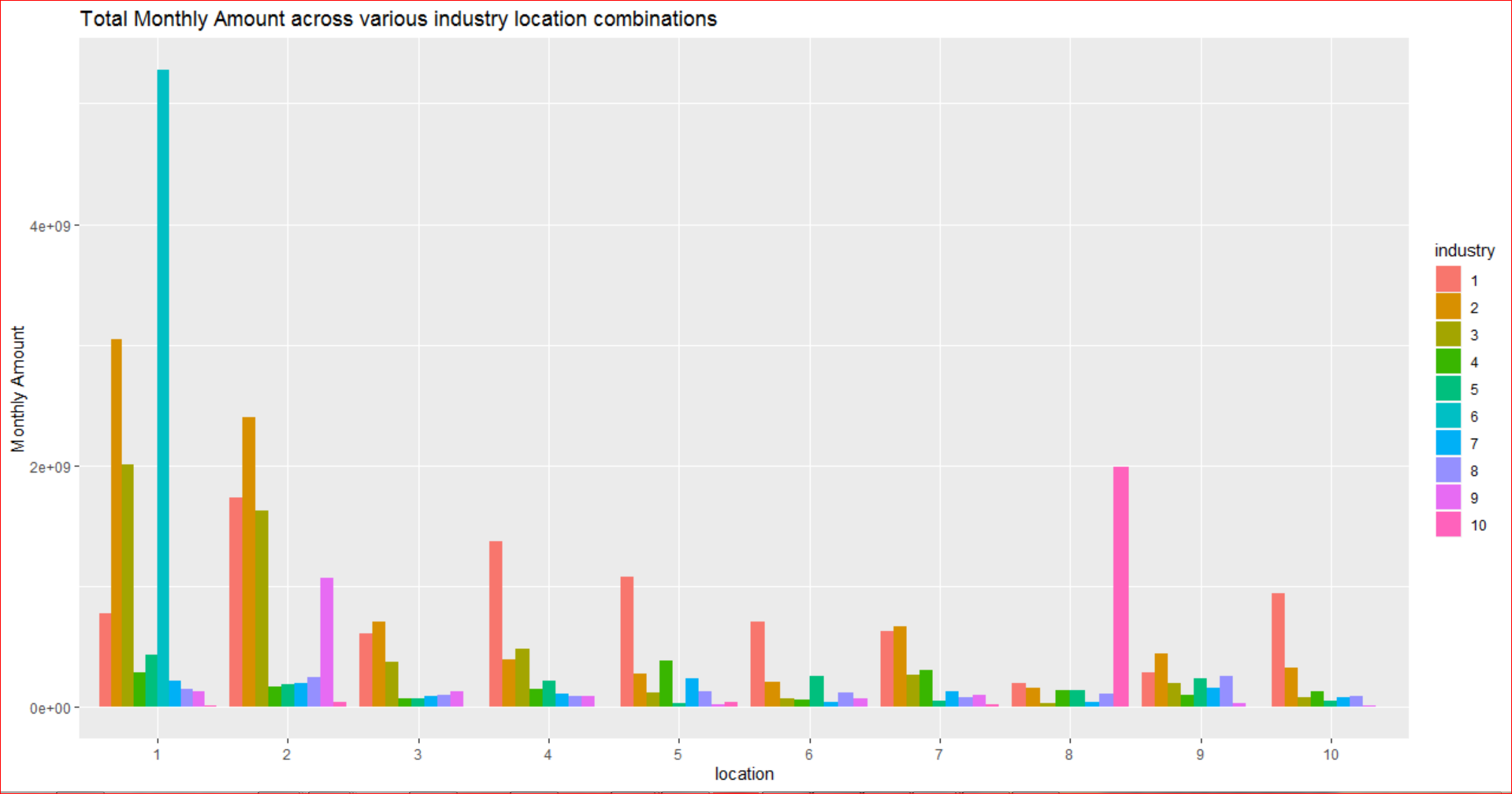


Fig 8 : Total monthly amount across various industry-location combinations

**Data Preparation :**

The provided data is currently at a date-customer\_id-industry-location level, which is not suitable for predicting the monthly amount for December 2016. Hence the data was aggregated to a industry-location-date level which calculates the mean of the monthly amount for each industry-location-date combination. The newly aggregated data contained 3886 rows of observations aggregated at industry-location-date level. However, some of the column variables were in incorrect format. Hence it was necessary to change a few of those variables to the required format for modelling(Fig 9).

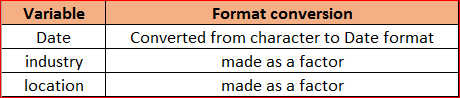


Figure 9 : format conversion of certain variables

The aggregated data was filtered for industry 1 location 1 combination and a subset of the dataset was formed containing 47 rows. Figure 10 shows a line plot of average monthly amount for industry 1 – location 1 with a gradually increasing trend.

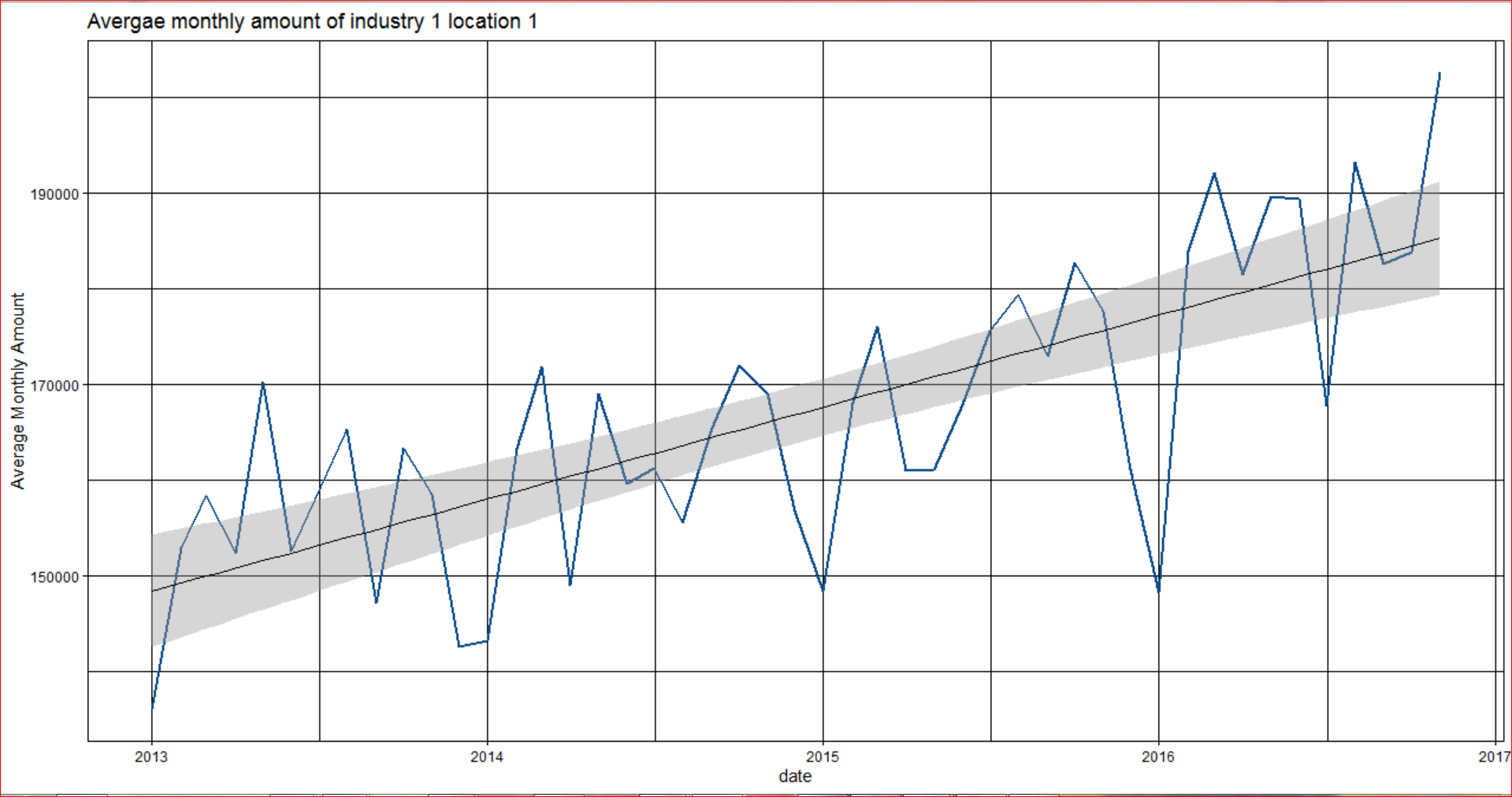


Figure 10 : Average monthly amount for industry 1 – location 1

Upon decomposing the above line plot we can observe that there are 2 components to it which are seasonality and trend(Fig 11). In figure 11 the mean monthly amount for Industry 1 location 1 has an increasing trend and a monthly seasonality can be observed(Fig 12 & 13).

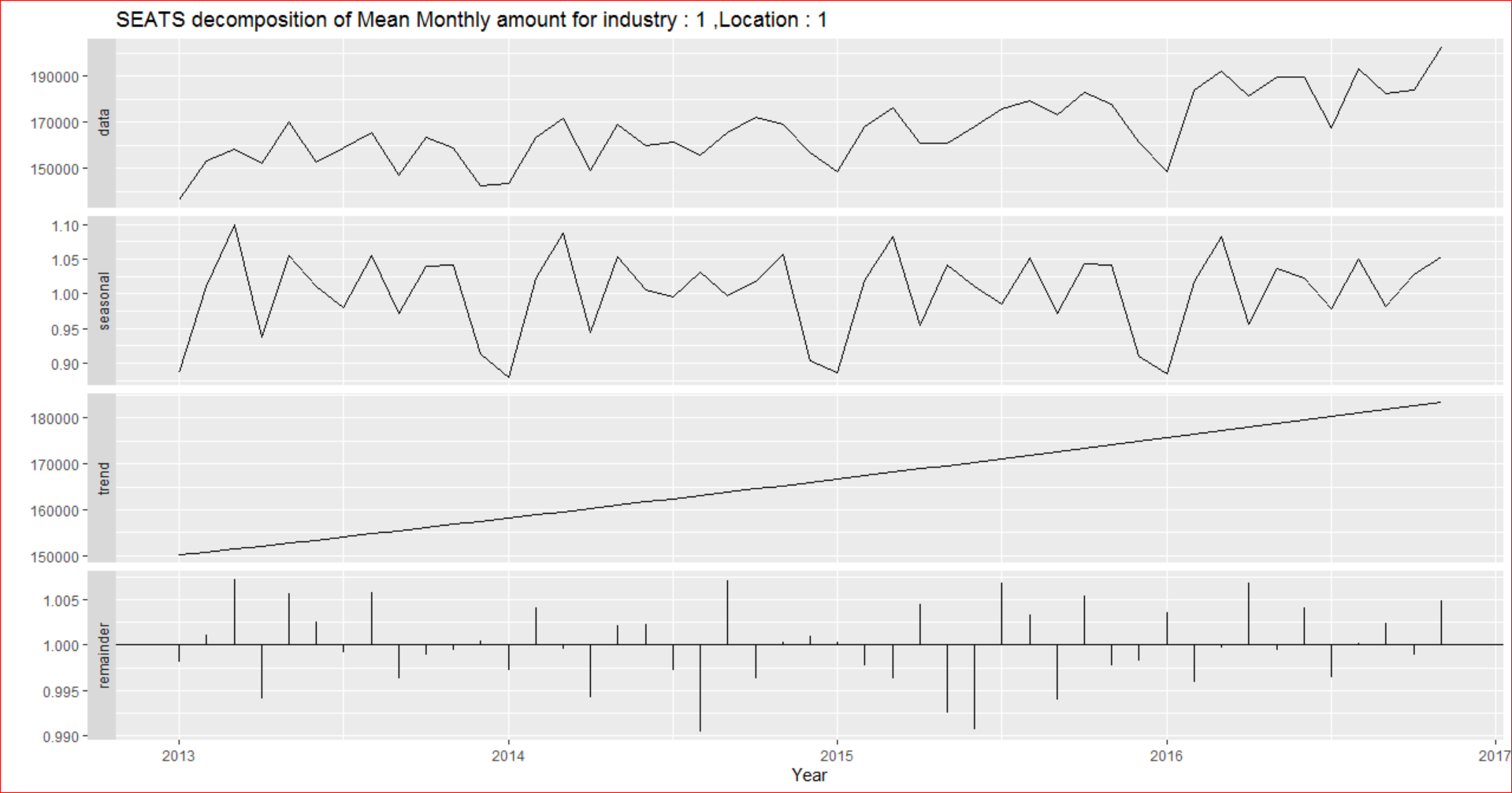


Figure 11: SEATS decomposition of mean monthly amount for industry 1 location 1

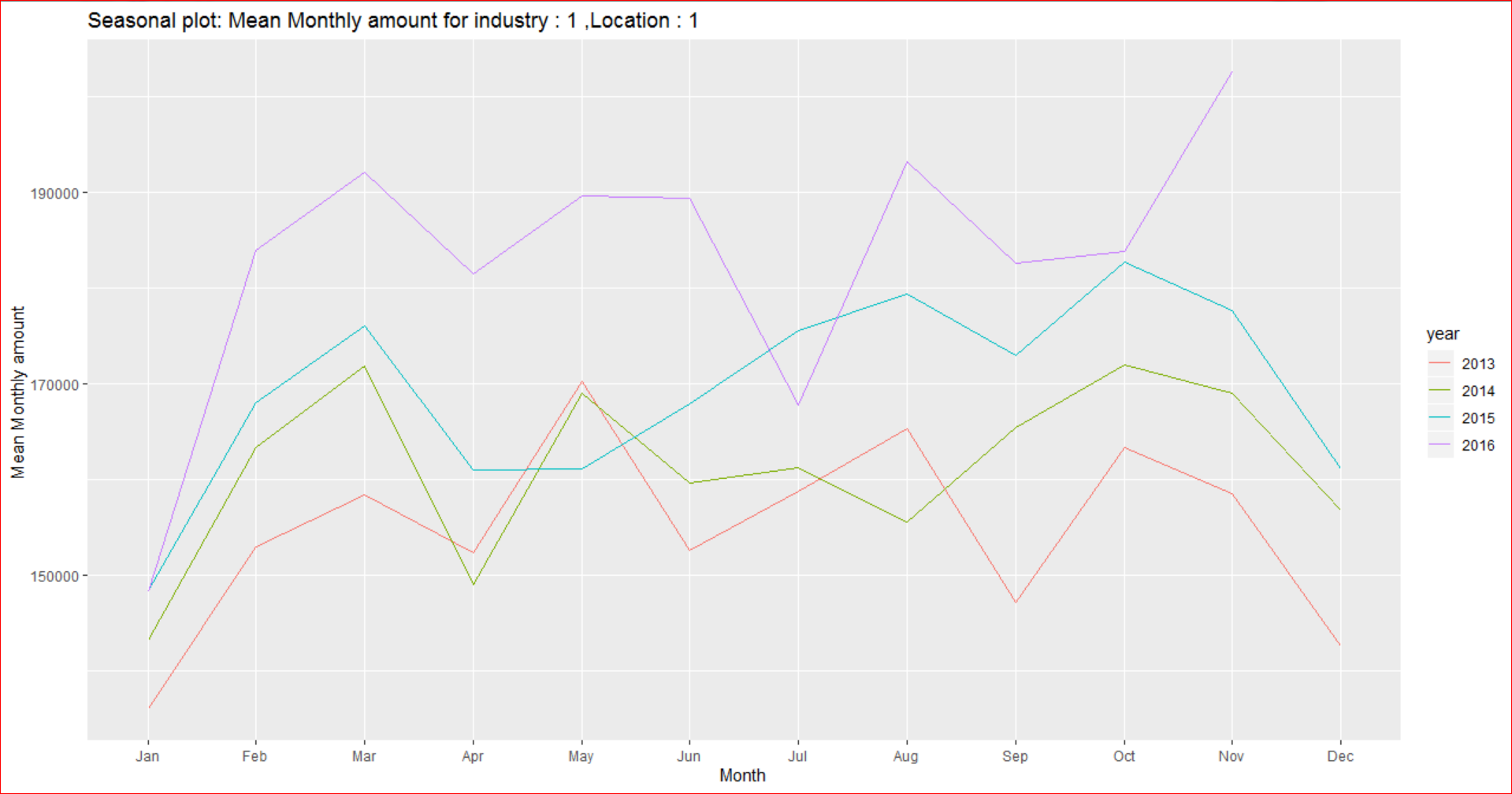


Figure 12 : Seasonal plot mean monthly amount for industry 1 location

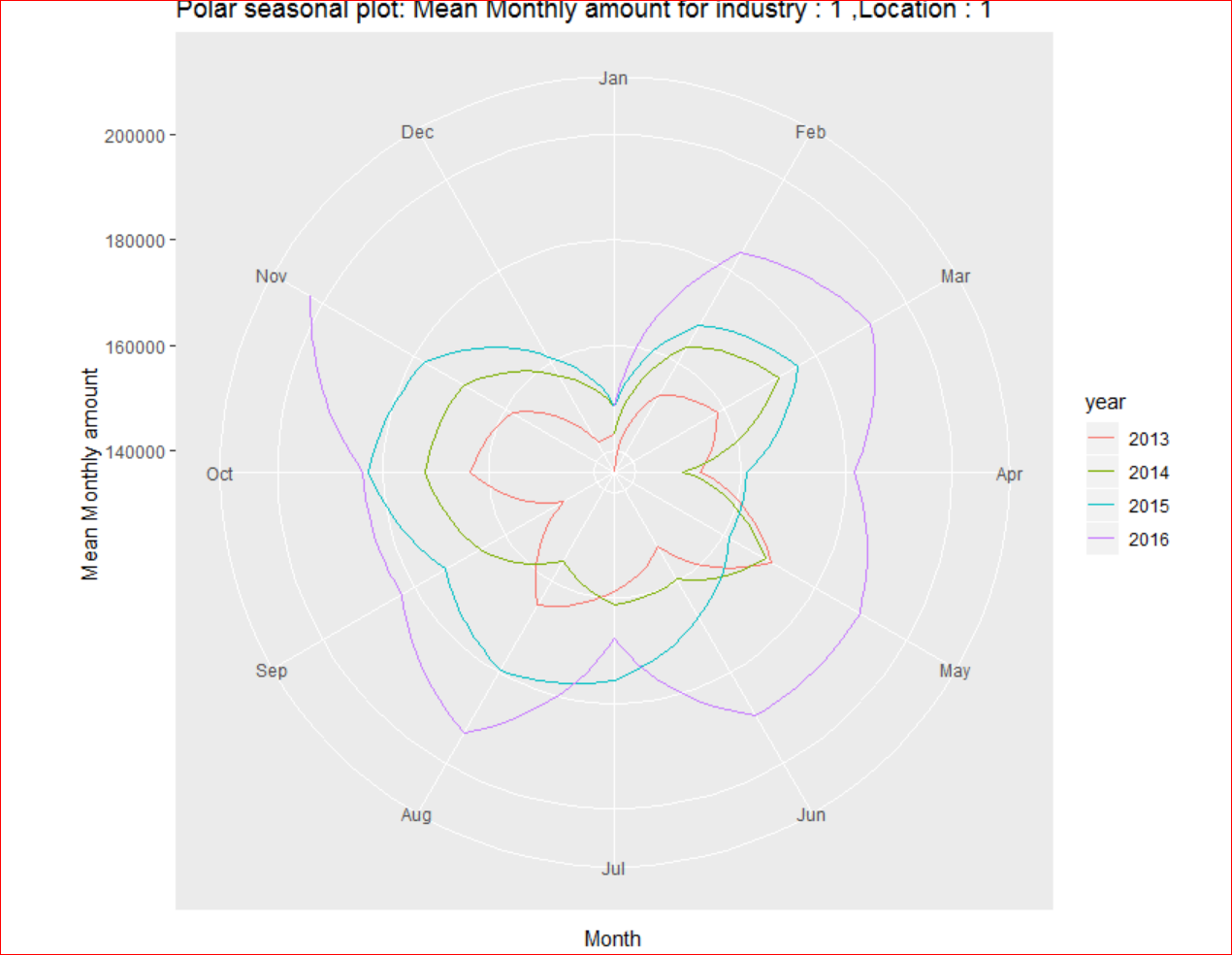


Figure 13 : Polar Seasonal plot mean monthly amount for industry 1 location

Hence it is important to capture these seasonality and trend of the time sequence as variables and use them as predictor variables for the model.

**Modelling :**

The dataset containing industry 1 location 1 values was further split into train and test set for modelling, in such a way that the train set accommodates the first 36 months ranging from 2013 to 2015 and the test set the remaining 11 months. This proportion of splitting the train & test set(36:11) was finalised after careful examination, so that the model could get an ample range of observations to train upon.

Considering that the target variable(monthly amount) is a numeric and a linear trend is observed between the date and monthly sale amount a linear regression model was chosen for the problem. In order to capture the seasonality and trend component of the data, the month and year was extracted from the date column and made available as separate predictor variables. The month was created to capture the seasonality component and the year for the trend component. The year was treated as a numeric to handle the increasing trend as it increased linearly with the monthly amount and the month was treated as a factor.

Upon training the model The Adjusted R-Square was observed to be 0.7457(Fig 14).

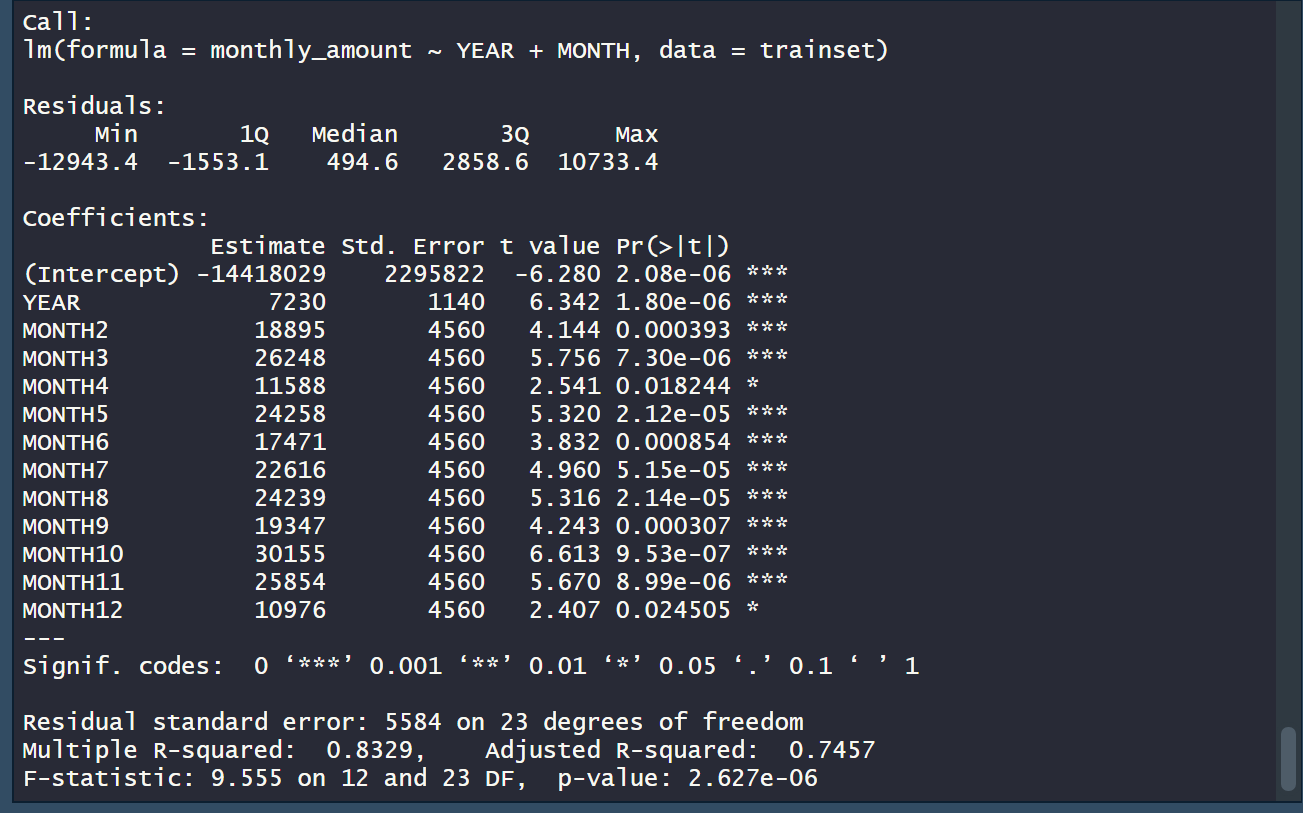


Figure 14 : Statistics of the linear regression model used for predicting mean monthly amount for industry 1 location 1.

**Evaluation :**

The goodness fit of a linear model is evaluated based on the RMSE parameter. Here the RMSE value is 11304.63. Figure 15 illustrates the difference between the actual and predicted values. Considering this model as the most appropriate, a prediction for December 2016 industry 1 location 1 was made and it gave the value of 168003.5. This value seems to be accurate as it follows the linear trend and is greater than the monthly amounts of December 2014 and 2015. Moreover, it also follows the December seasonality as the predicted value for December 2016 is lower than the actual value of November 2016.

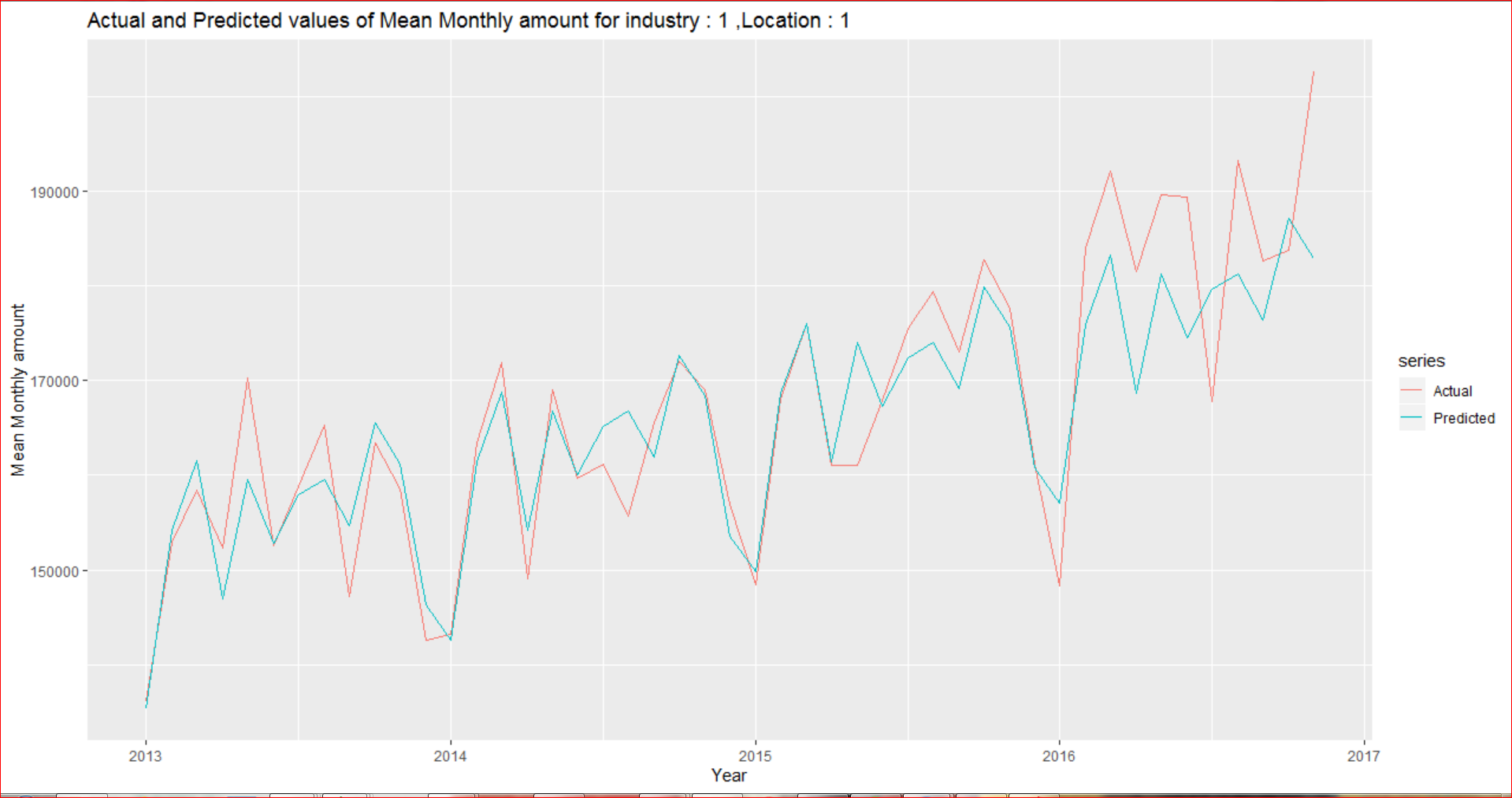


Figure 15 : Actual and predicted values of Mean monthly amount for industry 1 location 1

**Advanced Model Fitting :**

The model applied for industry 1 location 1 was used to make predictions for the other industry location combinations; programmatically using a for loop. However the below conditions were applied to filter out few industry location combinations in order to obtain accurate results :

* A minimum of 15 months of monthly amounts is required for each industry-location combo.

Upon applying these conditions a total of 84 industry-location combinations remained. The modelling process was iterated for these combinations and the predictions have been calculated(Refer Appendix).

Out of all the industry-location combinations to which the model was applied the modelling method gave worst results on the below mentioned Industry-location combinations(Fig 16) :

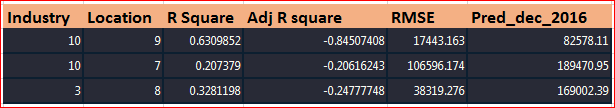


Figure 16 : Industry location combinations for which the model performed worst

The model performed poorly on the above combinations as we can see that the Adjusted R Squares are negative and the RMSE values are high for all of them. The reason for the poor performance was due to predictor variables from a subset of data can’t be used for all combinations in the data. Moreover, there were insufficient or very few observations in the training set which affected the models prediction accuracy.

**Appendix :**

R code :



Predictions on test set : ( predictions, the actual, the difference)



Evaluation measures on training data, testing data and predictions for December 2016 :

