**CE 472 – Computer Applications in Highway and Traffic Engineering**

**Homework 2 – Statistics with R**

1) The given ad and the police statistics states that most burglaries occur during the vacation period (time span = 3 months) i.e., between May and August (assuming from May 1st to August 1st) and which is 25% of the total burglaries. In order this to be true, all other 9 months must have a burglary rate less than 25/3 %.

Therefore,

Total burglary rate in a year < 25 + 9\*(25/3)

< 100%

Total burglary rate in a year must sum up to 100%. But the above calculations contradict that. Therefore, the statistics doesn’t prove that burglars go to work when other people go on vacation.

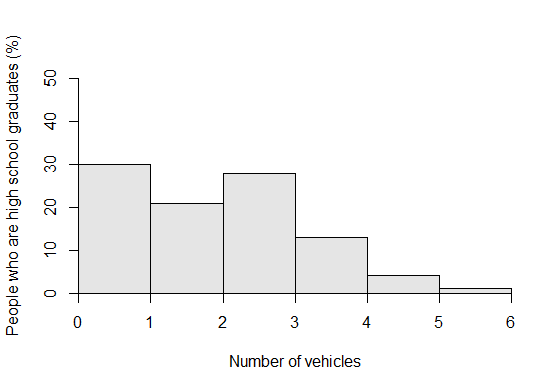
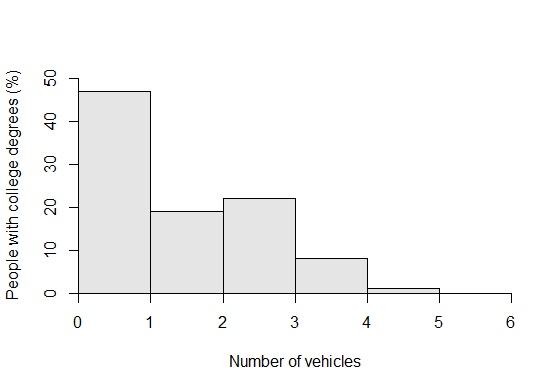
2)

|  |  |  |
| --- | --- | --- |
| Quartile | Income in 1973 | Income in 2004 |
| 25th | 5 | 30 |
| 50th | 10 | 60 |
| 75th | 15 | 100 |
| 100th | 50 | 200 |

NOTE: The below calculated values ae approximation based on the given histogram

As we can see from the above table, income of family has increase nearly 6 times for every quartile with the exception of 100th quartile. So we can conclude that family income went up by a factor of 4 over 30 years.

3.a) Number of vehicles owned is a discrete variable.

3.b)

3.c) Majority (nearly 50%) of the people among the college graduates are not able to afford a vehicle while only 70% of the high school graduates have a vehicle. This may be as a result of college graduates trying to pay off their student loan debts, so they are not able to afford a vehicle. This might be the same reason why % of people owning 1 is greater for high school graduates than college graduates (same goes for No. of vehicles = 2,3,4 and 5).

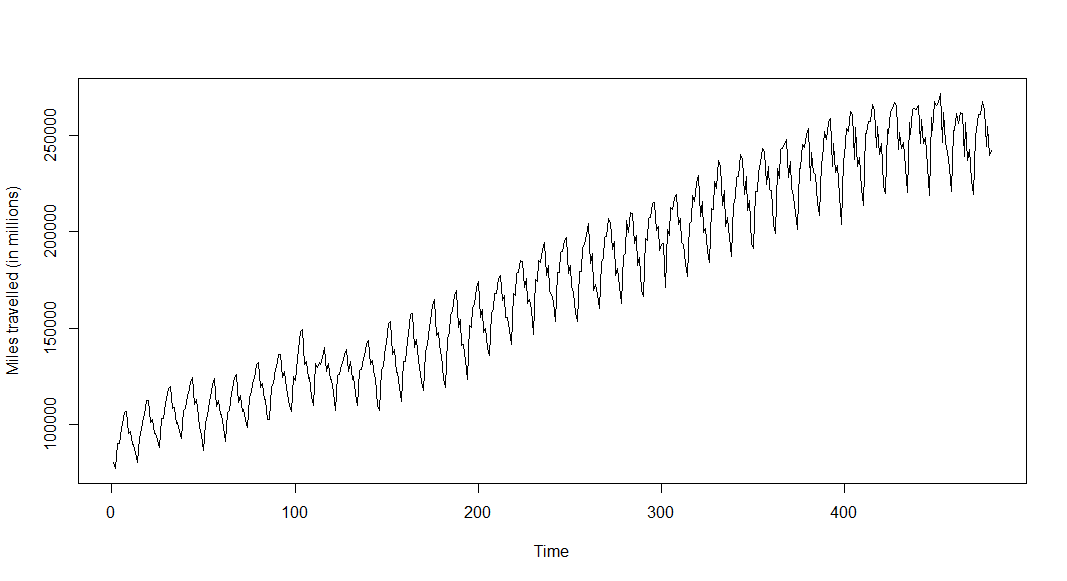
4.a) False. Those sections that liked their TA more did not do better on their final. The is slightly increasing for assistants rated < 3 and then decreases at a higher rate for assistants rated > 3.

4.b) True. There is almost no relationship between the section’s average rating of the assistant and the section’s average rating of the course. The course rating is close to 3 for all courses irrespective of the assistant’s rating.

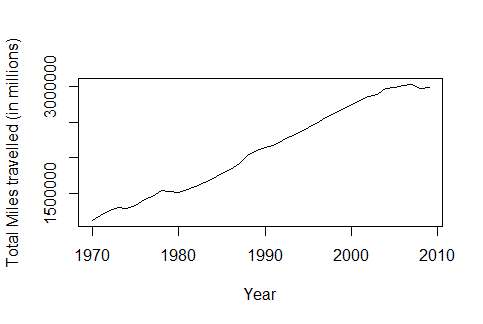
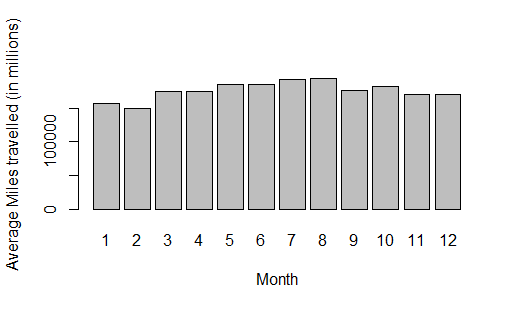
4.c) True. There is almost no relationship between the section’s average rating of the course and the section’s average score on the final. The course rating is close to 3 for all courses irrespective of the students’ average score on their finals.

5.a) Imported data successfully.

5.b) **Trend:** From the time series plot, we can say that the overall trend (of miles travelled) increases over the years and there is a periodicity for each year.



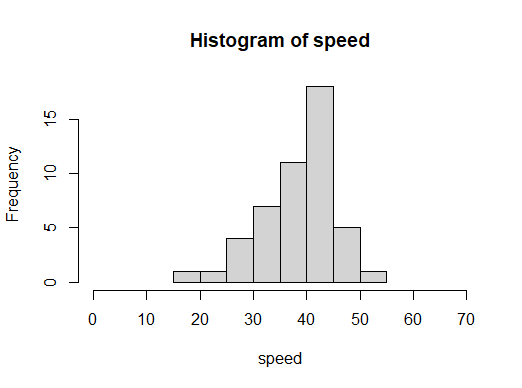
5.c)

5.d) Average gasoline prices for all categories from 1976-2009 have been imported from clipboard

5.e) In the given data for gasoline prices, there seems to some NA values (=10000000). Including the null values, I have removed the 13th month data which represents the average of gasoline prices over each year.

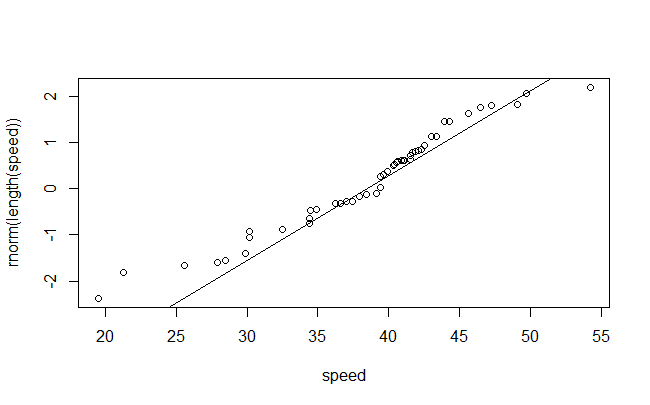
Correlation of gasoline prices v/s miles travelled (from 1976-2009) = 0.6747087. So, we can say that prices increase as miles travelled increases. This makes sense because, as travel length seems to increase over the years and as a result there is a high demand for gasoline. So gasoline prices go up.

6) Measures of central tendency and spread:

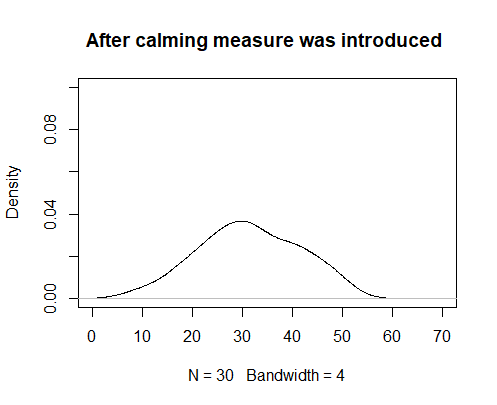
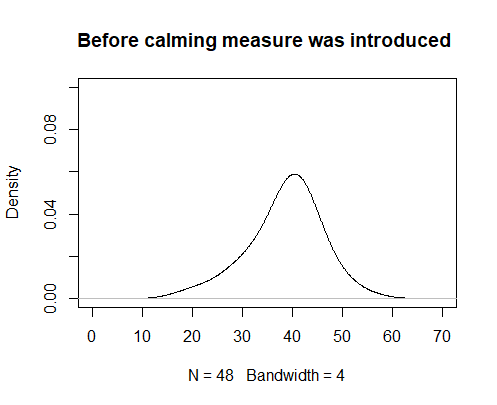
* Mean: 38.54
* Median: 40.09
* Mode: 30.2
* Standard deviation: 6.937628
* Range: (19.50, 54.20)
* Inter-quartile range: (34.80, 42.15)

85th percentile Speed: 43.8915 kmph

From the below qqplot we can say that the given speed data belongs to a normal distribution.



From the below plots, it is evident that there been a significant (around 10 kmph) reduction in speeds after a calming measure was introduced. Before the calming measure was introduced, most velocities were around 40 kmph. After the calming measure was introduced the majority of the speeds became 30 kmph.



By using descriptive statistics, we can also infer that even though the average velocity has decreased from 38.54 kmph to 31.33 kmph but the 85th percentile did not have significant decrease (43.891 kmph to 40.823 kmph) even after the calming measure was introduced.

**The same (check for significant reduction in speeds) can also be justified concretely using a t-test.**

**H0:** Speed after calming measure = Speed before calming measure

**H1:** Speed after calming measure < Speed before calming measure

By doing a Welch Two Sample t-test we get a p-value = 0.0005809. Since the p-value is less than 5%, we can safely reject the null hypothesis. As a result, we fail to reject the alternate hypothesis.

**Conclusion:** There has been a significant reduction in speeds

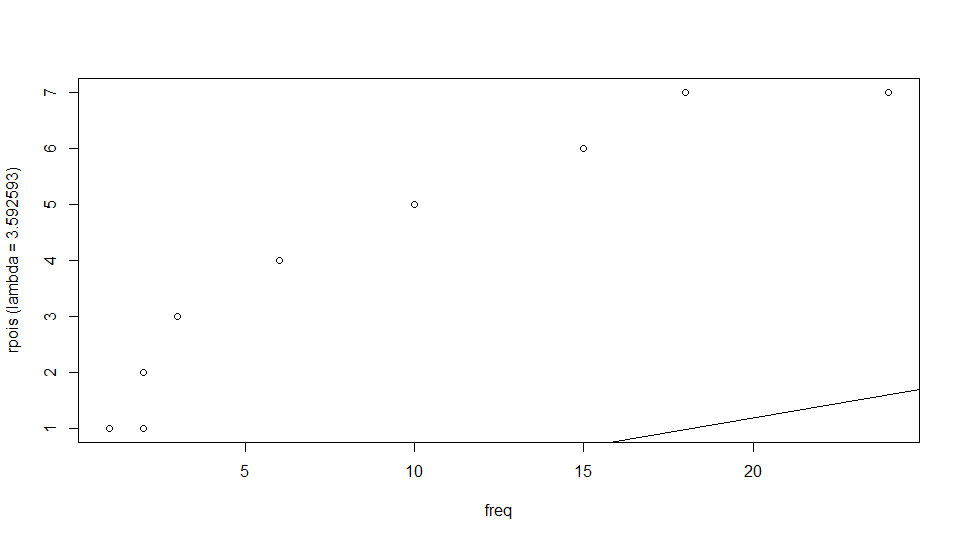
7) From the ANOVA test we get,

|  |  |  |
| --- | --- | --- |
| **Variable** | **p-value** | **Effect of categories (Rankings)** |
| Age | 3.74e-12 | 1 |
| Luggage | 0.000281 | 3 |
| Gender | 0.00643 | 4 |
| Age \* Luggage | 0.258 | No effect |
| Age \* Gender | 1.27e-05 | 2 |
| Luggage \* Gender | 0.311615 | No effect |

The effect of each variables has been ranked (1 represents highest effect) in the above table.

8) For the given frequency distribution, we get a lambda value = 3.592593.

Comparing the given frequency distribution with a Poisson distribution of lambda = 3.592593 using a qqplot (). The plot suggests that it is a very bad fit.

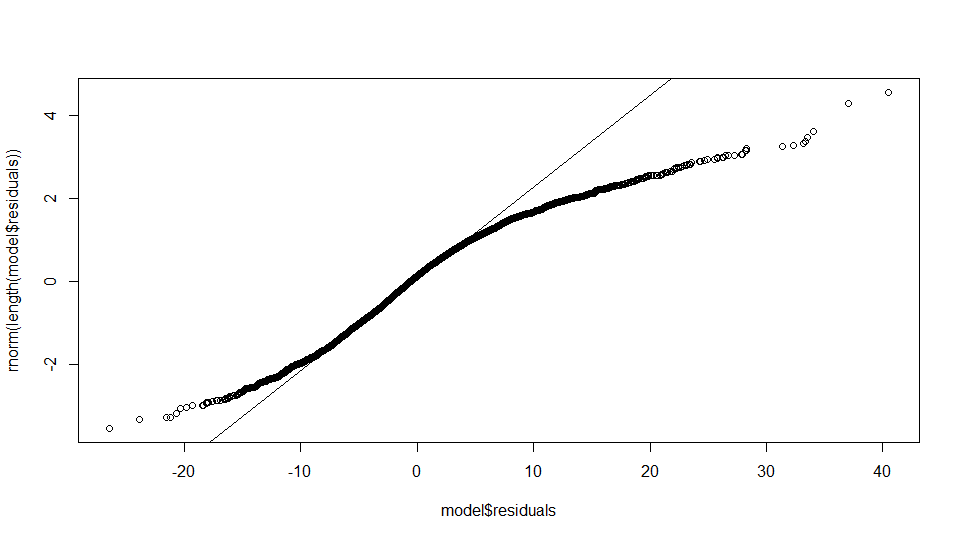


9) After doing appropriate exploration of the data I have chosen the following variables as independent variables for predicting CNTTDHH (No. trav day person trips made by HH). Due to the lack of clarity of what CNTTDHH comprises, I am assuming it includes public transports also.

* **HHSIZE and DRVRCNT:** Count of HH members and drivers in HH respectively. These two are used because “HHSIZE – DRVRCNT” gives us the no. of people in a HH that use public transport for their travels.
* **TRAVDAY:** Travel day (day of week) – Travel may vary with weekend and week days.
* **HHFAMINC:** Total HH income last 12 months – Income might be used to conclude if a HH can afford to travel.
* **HHNUMBIK:** No of full-sized bicycles.
* **RATIO.DRVEH (dummy variable):** Ratio of no. of drivers to that of no. of vehicles available – Using no. of driver and vehicles doesn’t directly affect CNTTDHH always. For example, if there are ‘10’ drivers and ‘0’ vehicles and vice-versa then it would result in zero travels. In cases where no. of vehicles (or) both vehicles and drivers are zero then we will get NA or Inf as the ratio. In such cases, I have replaced them with zero

Since all the independent variables I used above are categorical data, there will be no need for any dummy variables. Using dummy variables to account for non-linear effects makes sense only when our independent variable is contains continuous data.

From qqplot() and K-S test (p-value < 2.2e-16), we can conclude that errors are not normally distributed. We can’t use Shapiro’s test for this data. In order to use it we must have sample size between 3 and 5000.



**Remedies for non-normal distribution of residuals:**

All tests for normality show that the residuals are not normally distributed. It leads to fluctuations in the error of the model and it won’t be consistent across the full range of data. We can apply transforms on the data to make the residuals normally distributed, some of which are;

i) square root / cube root

ii) logarithm

iii) inverse