Bike Share Analysis

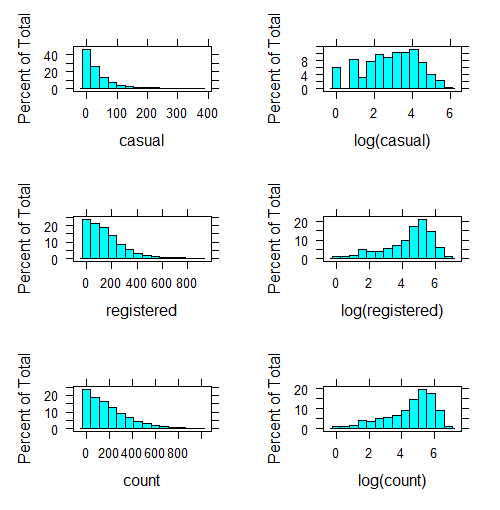
1. Spend some time reviewing the data and making necessary data transformations. Also read the last part of this question so that you are aware what we are looking for. First, briefly describe in your Word document what data transformations you made and why those transformations were necessary. (2 points)

That last question requires some feature engineering to answer the questions. To answer 5.a, 5.b and 5.e we need to convert weather, hour and workday variables into factor to get the effect of each kind of weather, each hour and each kind of day (Weekday or weekend) on each DV.

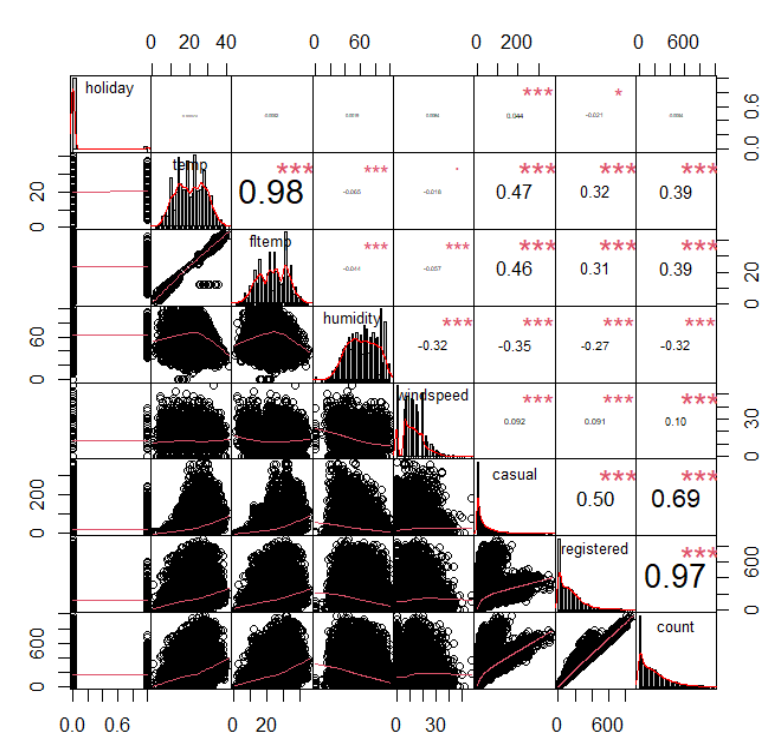
We need to split the date variable to get month and day to see the effect on day and month wise to answer 5.c and 5.d.

2. Present appropriate data visualizations and describe what you infer from these visualizations in your Word document. Do not go overboard with unnecessary visualizations as this will simply waste time and add unhelpful clutter. (1 point)

Let’s look at the distributions of the 3 independent variables here. (Casual, Registered and Total Count)



The distributions of all three DVs is right-skewed, and hence, OLS regression will not be suitable. The distributions of log-transformed transaction amount and count are close to normal, and therefore, more suited for OLS regression. We need to fit 3 models for each of the 3 DVs.



High correlation between temp and fltemp, We can use any one of them and drop the other. Here, let’s go with fltemp. High correlation between the DVs (0.7 and 0.98) and hence we should not use them as dependent in another model.

3. What variables do you think are pertinent to predicting the three dependent variables of interest. Create a table with these predictors, the sign of the expected effect, and your reasons justifying why you think these variables are appropriate. Also describe why the remaining variables are not important. If you think that some variables may have interaction or non-linear effects, justify that too in this predictor table. Remember that wrong choice of predictors will give you wrong interpretations. (3 points)

Since all three DVs are counts of rentals, we can use one predictor table.

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Effect** | **Rationale** |
| *DV: Casual, Registered, Count* | | |
| Weather | +/- | Rentals may be high on a clear day compared to a misty or a rainy day. |
| Fltemp | + | People may prefer to rent a bike on warmer days than colder or rainy days. |
| Humidity | +/- | High humidity can decrease bike rental counts due to discomfort and difficulty in riding in hot conditions. However, in cooler temperatures, it may increase demand as it can make the weather feel more pleasant and reduce the risk of dehydration. |
| Windspeed | +/- | People would not prefer to rent bikes on a windy day compared to non-windy days. |
| Hour | +/- | Rentals can be high during the early hours compared to late hours. |
| Year | +/- | The year can impact bike rentals in either direction due to changes in weather, economic conditions, trends in transportation and leisure activities, and location-specific factors. |
| Month | +/- | Some months like Spring, Summer and Fall months may see a rise and a decline in Winter or Rainy months. |
| Day | +/- | People might prefer to rent on weekends compared to weekdays. |
| Workday | +/- | Same reason as Day. |
| Excluded |  |  |
| Season | n/a | Seasonality trends in rentals is captured by month, day, and year. |
| Holiday | n/a | Correlation with workday == 0 |
| Temp | n/a | Correlation with fltemp |
| Date | n/a | Split into day, month, and year |

4. Run three models in R (one model for each dependent variable). Paste the R code for the three models and stargazer output showing the results of these three models. If you test multiple models, present only the "best" model for each DV. Also explain why you chose these models for your analysis and describe to what extent their assumptions are met. (2 points for choosing the right models + 1 point for model explanation + 1 point for assumption testing)

Model Building approach-

1. I tried to run OLS models for each of the DVs. The normality and homoscedasticity assumptions were violated in all the 3 models.
2. I tried to run poission models because it fits the count nature of the 3 DVs well. But the problem of overdispersion was quite significant.
3. To mitigate this, I tried running quasi-poisson models and negative binomial models. These models are robust to Overdispersion, and we did not have an excess zero problem in the data.
4. Multicollinearity in the quasi and nb models could be given a pass because it was observed only between day and workday variable, and we needed to include them both to answer the questions. Independence assumption failed in both the type of models.
5. We certainly need a time-series model like ARIMA which is designed to account for the correlation and dependence structure in time series data.
6. I’m going with the negative-binomial model to answer the questions. We need to note that it violates the assumption of Independence.

library(MASS)

casual\_nb = glm.nb(casual ~ workday + weather + fltemp + humidity + windspeed + hour + day + month + year, data = d)

registered\_nb = glm.nb(registered ~ workday + weather + fltemp + humidity + windspeed + hour + day + month + year, data = d)

count\_nb = glm.nb(count ~ workday + weather+ fltemp + humidity + windspeed + hour + day + month + year, data = d)

stargazer(casual\_nb, registered\_nb, count\_nb, type="text", single.row=TRUE)

Stargazer output-

==========================================================================

Dependent variable:

--------------------------------------------------------

casual registered count

(1) (2) (3)

--------------------------------------------------------------------------

workday1 -0.542\*\*\* (0.035) 0.013 (0.033) -0.089\*\*\* (0.032)

weather2 -0.088\*\*\* (0.014) -0.038\*\*\* (0.013) -0.046\*\*\* (0.012)

weather3 -0.641\*\*\* (0.025) -0.454\*\*\* (0.022) -0.476\*\*\* (0.021)

weather4 -0.311 (0.645) -0.248 (0.518) -0.194 (0.513)

fltemp 0.050\*\*\* (0.001) 0.017\*\*\* (0.001) 0.023\*\*\* (0.001)

humidity -0.004\*\*\* (0.0004) -0.002\*\*\* (0.0004) -0.003\*\*\* (0.0004)

windspeed -0.003\*\*\* (0.001) -0.002\*\*\* (0.001) -0.002\*\*\* (0.001)

hour1 -0.449\*\*\* (0.043) -0.465\*\*\* (0.036) -0.472\*\*\* (0.036)

hour2 -0.742\*\*\* (0.045) -0.870\*\*\* (0.037) -0.855\*\*\* (0.036)

hour3 -1.342\*\*\* (0.050) -1.557\*\*\* (0.039) -1.509\*\*\* (0.038)

hour4 -2.002\*\*\* (0.059) -2.109\*\*\* (0.041) -2.065\*\*\* (0.040)

hour5 -1.817\*\*\* (0.056) -0.825\*\*\* (0.037) -0.890\*\*\* (0.036)

hour6 -0.674\*\*\* (0.045) 0.572\*\*\* (0.035) 0.503\*\*\* (0.035)

hour7 0.338\*\*\* (0.041) 1.588\*\*\* (0.035) 1.518\*\*\* (0.035)

hour8 0.983\*\*\* (0.039) 2.122\*\*\* (0.035) 2.057\*\*\* (0.034)

hour9 1.138\*\*\* (0.039) 1.518\*\*\* (0.035) 1.505\*\*\* (0.035)

hour10 1.394\*\*\* (0.039) 1.064\*\*\* (0.035) 1.152\*\*\* (0.035)

hour11 1.577\*\*\* (0.039) 1.181\*\*\* (0.035) 1.288\*\*\* (0.035)

hour12 1.668\*\*\* (0.039) 1.390\*\*\* (0.036) 1.473\*\*\* (0.035)

hour13 1.680\*\*\* (0.039) 1.354\*\*\* (0.036) 1.453\*\*\* (0.035)

hour14 1.708\*\*\* (0.040) 1.249\*\*\* (0.036) 1.383\*\*\* (0.036)

hour15 1.712\*\*\* (0.040) 1.315\*\*\* (0.036) 1.433\*\*\* (0.036)

hour16 1.704\*\*\* (0.040) 1.620\*\*\* (0.036) 1.669\*\*\* (0.036)

hour17 1.771\*\*\* (0.039) 2.142\*\*\* (0.036) 2.122\*\*\* (0.035)

hour18 1.578\*\*\* (0.039) 2.090\*\*\* (0.036) 2.054\*\*\* (0.035)

hour19 1.369\*\*\* (0.039) 1.759\*\*\* (0.035) 1.736\*\*\* (0.035)

hour20 1.121\*\*\* (0.039) 1.440\*\*\* (0.035) 1.425\*\*\* (0.035)

hour21 0.935\*\*\* (0.039) 1.174\*\*\* (0.035) 1.165\*\*\* (0.035)

hour22 0.742\*\*\* (0.040) 0.914\*\*\* (0.035) 0.912\*\*\* (0.035)

hour23 0.427\*\*\* (0.040) 0.517\*\*\* (0.035) 0.522\*\*\* (0.035)

dayFri -0.042 (0.040) -0.011 (0.037) -0.046 (0.036)

dayMon -0.228\*\*\* (0.036) -0.135\*\*\* (0.034) -0.169\*\*\* (0.033)

daySat 0.059\*\*\* (0.020) 0.032\* (0.019) 0.025 (0.019)

dayThu -0.336\*\*\* (0.041) -0.037 (0.038) -0.105\*\*\* (0.037)

dayTue -0.374\*\*\* (0.041) -0.100\*\*\* (0.038) -0.165\*\*\* (0.037)

dayWed -0.402\*\*\* (0.040) -0.080\*\* (0.037) -0.152\*\*\* (0.037)

monthfeb 0.187\*\*\* (0.033) 0.149\*\*\* (0.026) 0.152\*\*\* (0.026)

monthmarch 0.849\*\*\* (0.033) 0.206\*\*\* (0.027) 0.274\*\*\* (0.027)

monthapril 1.175\*\*\* (0.034) 0.330\*\*\* (0.029) 0.439\*\*\* (0.029)

monthmay 1.291\*\*\* (0.038) 0.589\*\*\* (0.032) 0.665\*\*\* (0.032)

monthjun 1.119\*\*\* (0.041) 0.586\*\*\* (0.036) 0.628\*\*\* (0.035)

monthjuly 1.095\*\*\* (0.045) 0.503\*\*\* (0.039) 0.567\*\*\* (0.039)

monthaug 1.087\*\*\* (0.044) 0.536\*\*\* (0.038) 0.579\*\*\* (0.037)

monthsept 1.116\*\*\* (0.040) 0.616\*\*\* (0.035) 0.667\*\*\* (0.034)

monthoct 1.188\*\*\* (0.036) 0.699\*\*\* (0.031) 0.741\*\*\* (0.030)

monthnov 0.978\*\*\* (0.032) 0.667\*\*\* (0.027) 0.690\*\*\* (0.026)

monthdec 0.640\*\*\* (0.033) 0.680\*\*\* (0.027) 0.661\*\*\* (0.026)

year2012 0.344\*\*\* (0.011) 0.531\*\*\* (0.010) 0.499\*\*\* (0.010)

Constant 0.768\*\*\* (0.052) 2.804\*\*\* (0.045) 2.975\*\*\* (0.044)

--------------------------------------------------------------------------

Observations 10,886 10,886 10,886

Log Likelihood -38,848.180 -56,922.120 -58,887.760

theta 4.031\*\*\* (0.076) 3.837\*\*\* (0.057) 3.904\*\*\* (0.057)

Akaike Inf. Crit. 77,794.360 113,942.200 117,873.500

==========================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

> levels(d$weather)

1. What is the effect of weather on total, registered, and casual bike rentals?

|  |  |  |
| --- | --- | --- |
| Total Count | Registered | Casual |
| Compared to clear days, the total rentals count for Misty/Cloudy days, Light snow/rainy days and Heavy snow/rainy days decreases by 4.6 %, 47.6% and 19.4 % respectively when all other predictors are held constant. | Compared to clear days, the registered rental counts for Misty/Cloudy days, Light snow/rainy days and Heavy snow/rainy days decreases by 3.8 %, 45.4 % and 25 % respectively when all other predictors are held constant. | Compared to clear days, the casual rental counts for Misty/Cloudy days, Light snow/rainy days and Heavy snow/rainy days decreases by 9 %, 64% and 31 % respectively when all other predictors are held constant. |

1. Are rentals of total, registered, and casual bike higher during weekends than during weekdays? By how much? Note: Holidays (e.g., Independence Day) are not necessarily weekends.

|  |  |  |
| --- | --- | --- |
| Total Count | Registered | Casual |
| Compared to the weekend, the registered casual rental counts are 9% lower on the weekdays when all other predictors are held constant. | Compared to the weekend, the registered rental counts are 1.3 % higher on the weekdays when all other predictors are held constant. | Compared to the weekend, the total rental counts are 54 % lower on weekdays when all the other predictors are held constant. |

1. Which month of year has the highest and lowest counts of total, registered, and casual bike rentals? What is the difference in rental count between these two months (with highest and lowest counts)?

|  |  |  |
| --- | --- | --- |
| Total Count | Registered | Casual |
| October has the highest total count for rentals and January is the lowest with the difference in rental count being 74 %. | October has the highest registered count for rentals and January is the lowest with the difference in registered rental count being 70 %. | October has the highest casual rental count and January is the lowest with the difference being 118 %. |

1. Which day of week has the highest and lowest counts of total, registered, and casual bike rentals? What is the difference in rental count between these two days?

|  |  |  |
| --- | --- | --- |
| Total Count | Registered | Casual |
| The highest total rentals count is registered on Saturdays and lowest on Mondays with the difference being 19 %. | The highest registered rentals count is registered on Saturdays and lowest on Mondays with the difference being 16.7 %. | The highest casual rentals count is registered on Saturdays and lowest on Wednesdays with the difference being 10 %. |

1. Which hour of day has the highest and lowest counts of total, registered, and casual bike rentals? What is the difference in rental count between these two hours?

|  |  |  |
| --- | --- | --- |
| Total Count | Registered | Casual |
| The highest hour of the day for total rental counts is hour 17 and the lowest is hour 4 with the difference being 4 %. | The highest hour of the day for registered rental counts is hour 17 and the lowest is hour 4 with the difference being 4.2 %. | The highest hour of the day for casual rental counts is hour 17 and the lowest is hour 4 with the difference being 3.7 %. |