The predictability and regularity of global UFO sighting durations

1. Introduction

From their depictions in conspiracy thrillers such as the X-Files to those in silly cartoons such as Marvin the Martian, UFOs and other such visitors from another world have long been a prominent part of popular culture. Indeed, man has long wondered if he is not alone in the vast emptiness of space that surround him. It is no surprise, then, that UFO sightings are a relatively common phenomena. Of particular interest is the duration of said sightings. It is often the case that the stories that endure the longest, the Roswells and the Lubbock Lights of modern mythos, are the stories with the most detail. Typically, this detail may only be attained if sightings are of sufficient duration and as such, it is often the case that sightings with longer durations have a longer, and greater impact.

In this study, I will be attempting to determine and hopefully explain some of the determinants of sighting duration. Particularly interesting is the day of the week during which the sighting occurred. The primary prediction would be that sightings that occur later in the week, peaking on weekends, would have longer durations.

In order to analyze the factors that determined the duration of a UFO sighting, I looked at data accumulated by the National UFO Reporting Center, collected prior to 2013. The dataset represents over 80,000 individual UFO reportings as old as 1943. Data points without a valid geographic location were excluded. Using a multiple regression model, I can explore the factors of sighting duration and determine which factors have the greatest influence.

The final model satisfies all of the assumptions made during a ridge regression. Outliers are certainly present however by the very nature of the data it is difficult to conclude when an outlier should be excluded. The data is almost entirely self-reported and the sources of the data are often less than reliable, increasing the challenge in drawing any reasonable relationships between data. The results from the residual plot and histogram of standardized residuals suggests that the data does not need to be further transformed.

The results suggest that there is indeed a strong correlation between the day of the week and the duration of a UFO sighting. Every increase in the day of the week corresponds with an average increase of 1356.8424102 seconds. This result is very much in line with the initial prediction of the possible effect that day of the week has on sighting duration. Though it is possible that the comments have an affect on the duration of a UFO sighting as more positive comments may correlate with longer reported durations, such an analysis would require a sentiment analysis to be performed on the set of comments which is outside the scope of this project. It was initially predicted that there would be a strong relationship between the shape of the UFO and the reported duration of the encounter. Arguably this may be related to the fact that some of the shapes more strongly resemble possible experimental aircraft than others. The final tested parameter was the location of the object.

2. Data

For the purposes of this study I used a dataset scraped from National UFO Reporting Center's database of UFO sightings found on kaggle. The dataset has been sanitized and invalid values have been dropped. As the purpose of the study was to determine the causes and determinants of UFO sighting duration, the dependent variable was the duration of the sighting in seconds (*duration*). For the independent variables I included three general groups of data, all with discrete values. The only value that was not encoded in any way was the day of the week variable, (*weekday*) which was extracted from the given date of the report. The dataset also supplied the estimated longitude and latitude of the report. The longitude and the latitude may be interpreted as a two dimensional matrix of points with the latitude representing the y coordinates and the longitude representing the x coordinates. I applied a row major order transformation in order to flatten the matrix into a single dimensional vector. I then used a sort of reverse row major order grouping method to place each of the coordinates into thirty two distinct geographic sectors where the longitude is split into eight sectors and the latitude is split into four sectors. Out of these thirty two sectors, only fourteen had any values, sectors 4, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, and 23 (sector*_4, sector_11, sector_12, sector_13, sector_14, sector_15, sector_16, sector_17, sector_18, sector_19, sector_20, sector_21, sector_22,* and *sector_23)*. This is not surprising as these values correspond to coordinates closer to the central regions of the map whereas many of the excluded sectors represent areas in geographic extremes. dataset also provided the city, state, and country of the report however these are trivially similar to the longitude and latitude and a high correlation is obvious. In this way, these independent variables may be interpreted as interaction terms between the longitude and latitude. The final independent variables are the group of dummy variables representing the shape (*'changed', 'changing', 'chevron', 'cigar', 'circle', 'cone', 'crescent', 'cross', 'cylinder', 'delta', 'diamond', 'disk', 'dome', 'egg', 'fireball', 'flare', 'flash', 'formation', 'hexagon', 'light', 'other', 'oval', 'pyramid', 'rectangle', 'round', 'sphere', 'teardrop', 'triangle',* and *'unknown'*). Of note is the distinction between the 'changed' and 'changing' shapes. 'Changed' means that the UFO changed once or more during the sighting but at some points held a consistent shape whereas 'changing' means that the UFO was constantly shifting. All told, there are fourteen distinct sectors and twenty nine distinct shapes.

Table 1 contains the descriptive statistics for a selection of my data. The average day of the week for a sighting was 3.15 meaning that it would have been between Thursday and Friday. For the most part, the dummy variables of the various sector have extremely small means, suggesting that there are not many significant concentrated areas of sighting. This is less consistent with the results for shape, where shapes such as 'circle' having higher means, suggesting that these shapes are more common than others such as 'changed'.

Though using a ridge regression greatly mitigates the effects of serious multicollinearity between independent variables, a correlation table is included in Table 2. The correlation table is displayed as a heatmap due to the large number of features. Even with the lack of distinct float values, it is clear that no serious correlation exists between variables. Though within the dummy variable sets this should be clear, it is surprising that there exists no correlation between location and shape as it is not unreasonable to suspect that certain geographic areas would have a tendency for certain shapes.

3. Regression Analysis

In order to create an appropriate model to estimate the duration of UFO sightings, I utilized a ridge regression. A ridge regression was chosen over an ordinary least squares regression as ridge regression greatly mitigates the negative effects of serious multicollinearity. Perhaps most significant, is that increases in multicollinearity resultant from the presence of outliers will have a greatly reduced impact on the accuracy of the final model. Indeed, this is especially important for this particular dataset as outliers can not be reliably be removed as any particular UFO sighting cannot generally be said to have any more credibility than another. Thus the resulting regression model is:

Duration = α + β1weekday + β2sector_4 + β3sector_11 +β4sector_12 + β5sector_13 + β6sector_14 + β7sector_15 + β8sector_16 + β9sector_17 + β10sector_18 + β11sector_19 + β12sector_20 + β13sector_21 + β14sector_22 + β15sector_23 + β16changed + β17changing + β18chevron + β19cigar + β20circle + β21cone + β22crescent + β23cross + β24cylinder + β25delta + β26diamond + β27disk + β28dome + β29egg + β30fireball + β31flare + β32flash + β32formation + β33hexagon + β34light + β35other + β36oval + β37pyramid + β38rectangle + β39round + β40sphere + β41teardrop + β42triangle + β43unknown + ε

As previously discussed, one of the major factors which I suspect have an impact on the duration of a UFO sighting is the day of the week on which the sighting occurred. Sightings which occur earlier in the week on working days are likely to be shorter as people tend to be busier and will have less time to either be in settings which are conducive to sighting a UFO or will have little time to pay attention to mild curiosities in the sky. Later in the week, particularly Friday through Saturday, which are encoded as 4-6, people will be out of their workplaces and are more likely to be in more remote areas with clearer skies. Furthermore, because they are not working, they may have more time to investigate anything even mildly interesting that they see in the sky.

Another important set of independent variables are those representing the various geographic sectors in which the sighting occurred. I suspect that those closer to densely populated areas are less likely to report a UFO sighting as urban light pollution would obscure much of the sky and research groups are less likely to test their aircraft near these population centers. What few reports are generated from these areas are unlikely to be very long in duration for the same reasons that they may not be very frequent. On the other hand, those in more remote areas who are also close to air bases or aircraft research groups may report a greater number of UFO sightings and report sightings with longer durations.

The final set of independent variables are those representing the shape of the reported UFO. I believe that the reported shape may have a significant impact on the sighting duration as shapes which more closely resemble certain types of experimental aircraft or natural atmospheric phenomena may have longer durations whereas those with unknown shapes or shapes more closely resembling high speed experimental aircraft may have shorter durations.

The results of the ridge regression model are display in Table 3. It may come as little surprise that the R-squared value for the model is 0.0007025. Additionally, the overall significance of the model is abysmally far from acceptable. Of all the independent variables, ultimately only one is even close to significant, the day of the week on which the sighting occurred. None of the dummy variables have any significance at the five percent level and in fact, several of them have no significance at any level with p values of 1.000.

The histogram of standardized residuals for the full model is relatively normal and suggests that the error terms generally follow a normal distribution. There is no heteroskedasticity evident from the residual plot as the residuals generally follow closely to the predicted variables with the exception of a relatively insignificant of outliers. The correlation matrix has already shown that there is no correlation between the independent variables and even if there were, the ridge regression method should reduce the impact of any such correlation.

To create the final, reduced model I dropped all of the dummy variables, leaving only the day of the week. Though there is no change in the R-Squared and Adjusted R-Squared values, there is an extremely significant increase in the overall F test for validity and a vast improvement in the p-value for the only independent variable maintained, the day of the week. This absence of change suggests that the previously included independent variables had no significant effect on the regression. The only issue is that the residuals for the reduced model are slightly less consistent than the residuals of the full model though they do not have enough variation to constitute heteroskedasticity. The calculated partial F test statistic was 0.0882, significantly less than the significant F value at $\alpha=0.05$, 0.67. Thus there is not enough evidence to reject the null hypothesis and we may not conclude that at least one of the dropped variables was not equal to zero. The remaining independent variable has a p-value far below 5% and as no assumptions were violated, the final regression model should be:

$$\textbf{Duration} = \alpha + \beta 1\textbf{weekday} + \varepsilon$$

4. Empirical Results

The results of the final regression model as displayed in Table 4 suggest that each increase in the day of the week corresponds to an increase of 2085.7 seconds in duration of UFO sighting. Given that the average duration of a sighting is approximately 8000 seconds this is quite a significant increase of almost 25%.

Indeed, as earlier predicted, the day of the week has the most significant impact on the duration of the sighting. It is interesting that the shape and location of the sighting appear to have no significant effect on the duration. Though it may be easy to dismiss this observation as attempting to apply too much of a logical approach to a dataset which is ultimately relatively unreliable, the degree to which these independent variables had no impact is surprising.

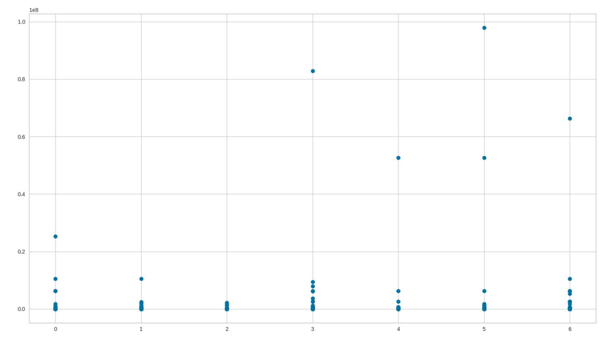
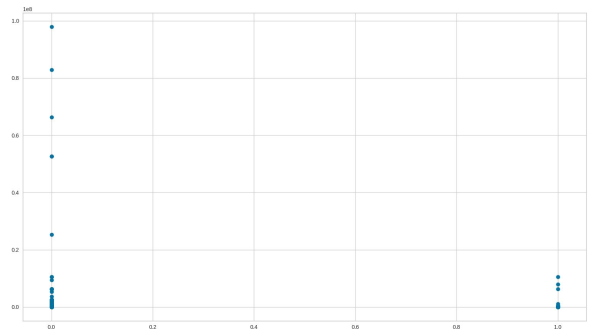
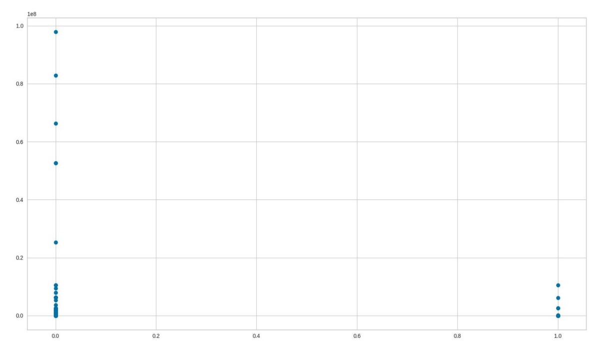
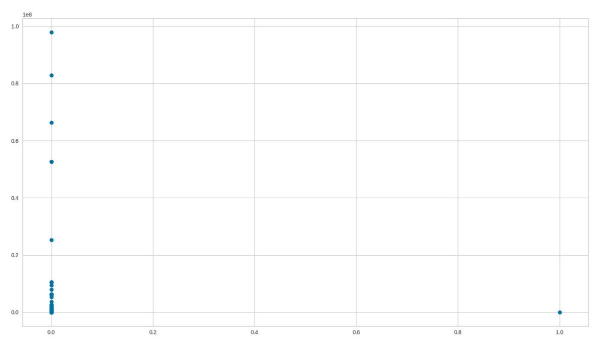
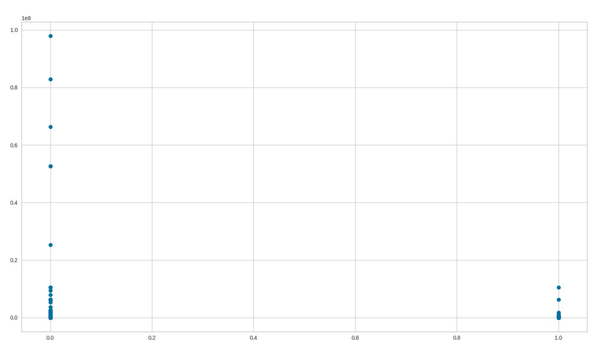
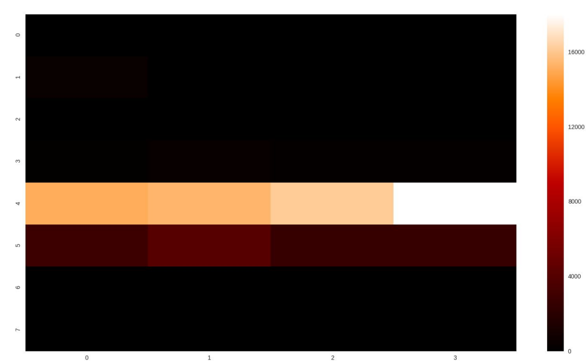
5. Summary and Discussion

The study attempted to discern the factors which determined the duration of UFO sightings reported around the world. I was interesting in determining whether a number of circumstances surrounding each sighting had a significant effect on the duration of the sighting and whether these circumstances could be used to predict future sighting durations. In the end, I found that the only truly important and significant factor of the sighting duration is the day of the week of the sighting. Indeed, an increase in day of the week corresponded to an almost 2000 second increase.

One minor issue from the study methodology was the lack of a separation between a training and a predicting dataset. In keeping with the methods used in the course, the same data was used for both the training and the scoring though in typical applications of the utilized data analysis tools, a random sample would have been selected from the dataset to be partitioned off as a prediction dataset while the remaining data would be used as a training set.

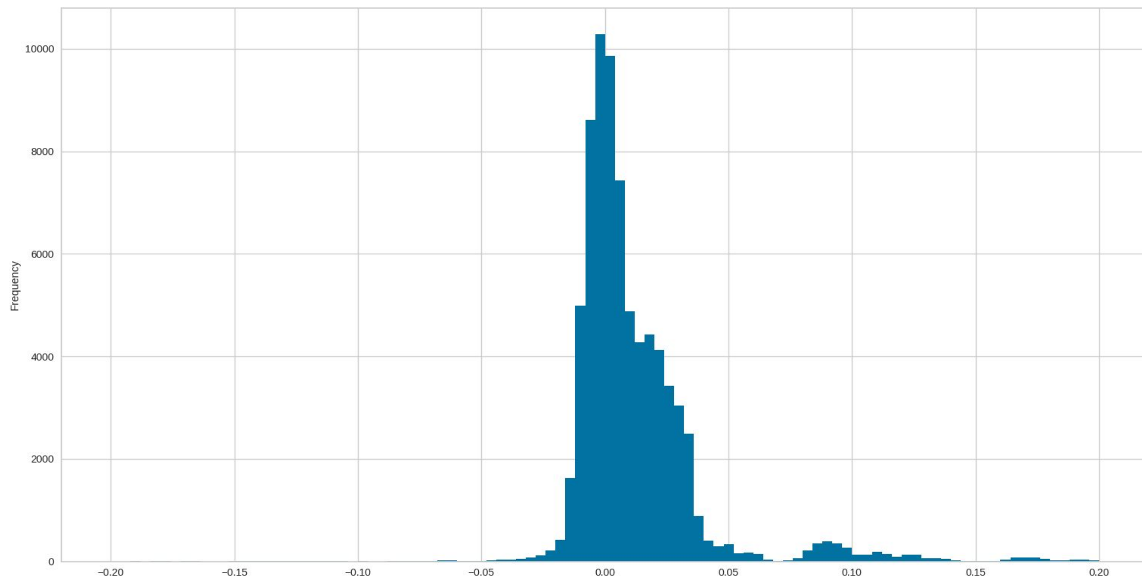
Perhaps the greatest issue in this study is that these events are, by their very nature, unpredictable. Though we may postulate what goes into each sighting and attempt to dissect each case, the data will be ultimately extremely biased. The vast majority of sightings are almost certainly unreported and are easily dismissed. Most sightings are likely natural atmospheric phenomena or simply just unfamiliar aircraft. It is possible, that with sufficiently advanced sentiment analysis techniques, the comments and reported description of each encounter may be used to see if there is an correlation between the positivity of a comment and the duration. This, however, would be relatively difficult as the tone, content, and language of the comments in the reports differ quite significantly. Ultimately, only one thing is certain from the results of this study, that aliens prefer to visit on weekends.

Figure 1. Scatter plots of duration vs independent variables and heatmap of sector frequency

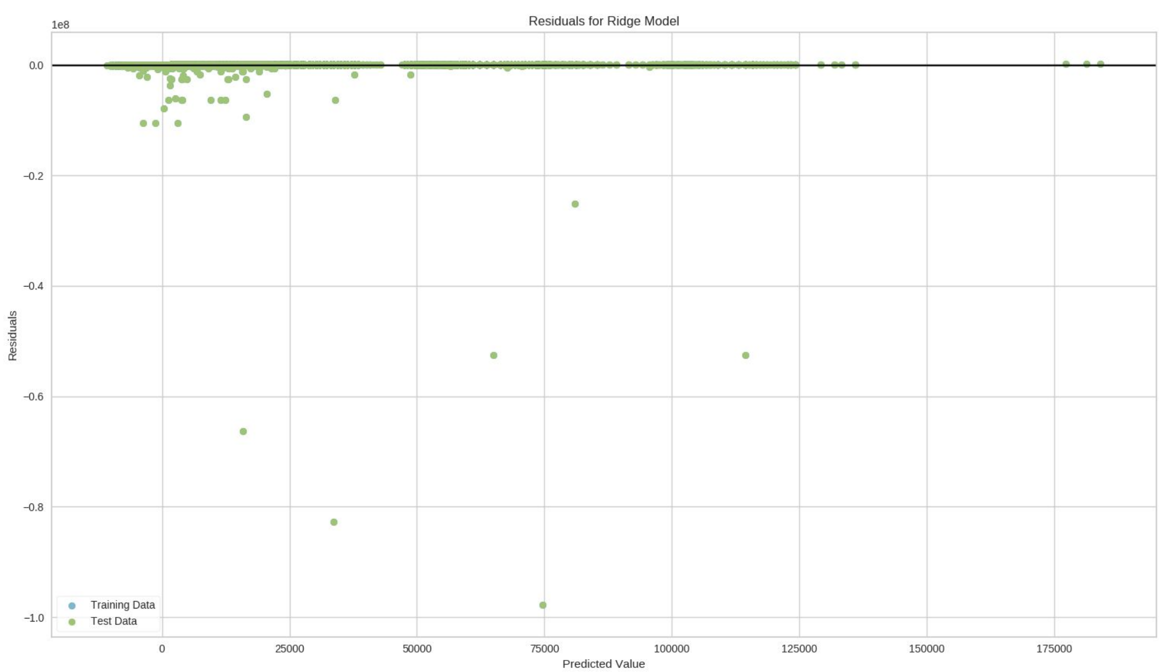
	
Day of Week Plot	Circle Plot
	
Fireball Plot	Pyramid Plot
	
Unknown Plot	Sector Heatmap

Note: Only a selection of shape plots have been included, they are generally the same

Figure 2. Histogram of Standardized Residuals and Residuals vs. Predicted duration of full model

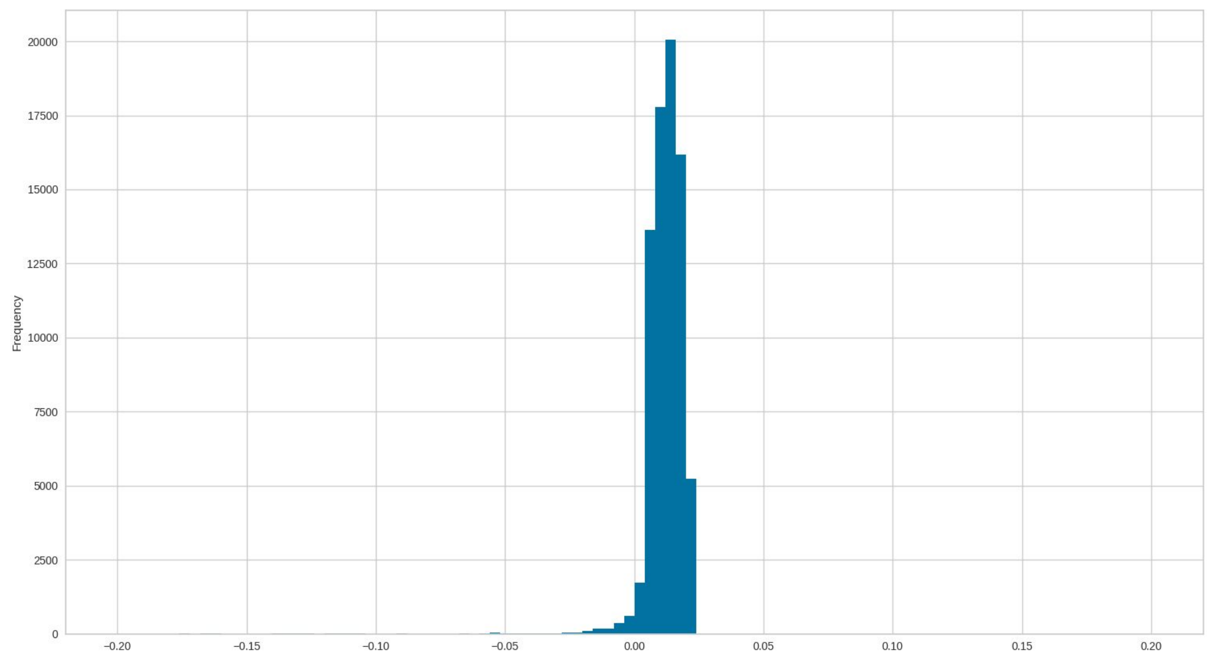


Standardized Residuals Histogram of Full Model

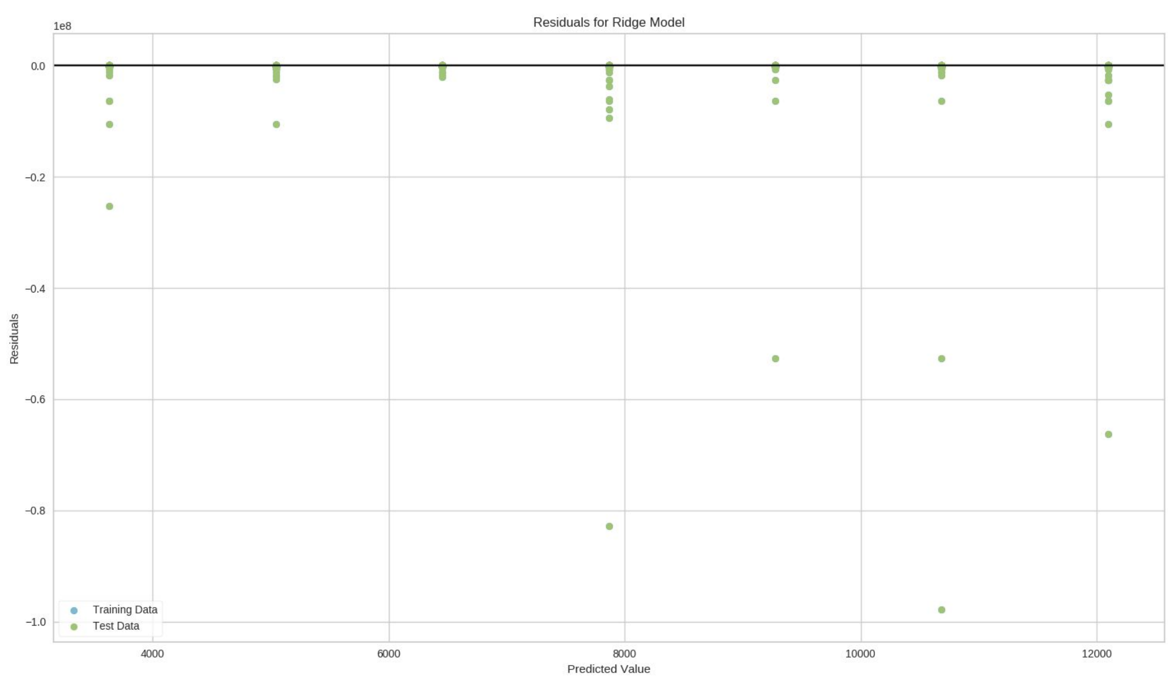


Residuals vs. Predicted durations of Full Model

Figure 3. Histogram of Standardized Residuals and Residuals vs. Predicted duration of reduced model



Standardized Residuals Histogram of Reduced Model



Residuals vs. Predicted durations of Full Model

Table 1. Descriptive Statistics

	changed	changing	chevron	cigar	circle
count	78398.000000	78398.000000	78398.000000	78398.000000	78398.000000
mean	0.000013	0.025026	0.012143	0.026238	0.097031
std	0.003571	0.156205	0.109526	0.159843	0.296001
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000

Shapes (incomplete)

count	78398.000000
mean	3.153945
std	1.985042
min	0.000000
25%	1.000000
50%	3.000000
75%	5.000000
max	6.000000
Name: datetime, dtype: float	

Day of the week

	sector_4	sector_11	sector_12	sector_13	sector_14
count	78398.000000	78398.000000	78398.000000	78398.000000	78398.000000
mean	0.006276	0.000013	0.002589	0.004528	0.003036
std	0.078971	0.003571	0.050820	0.067140	0.055015
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000

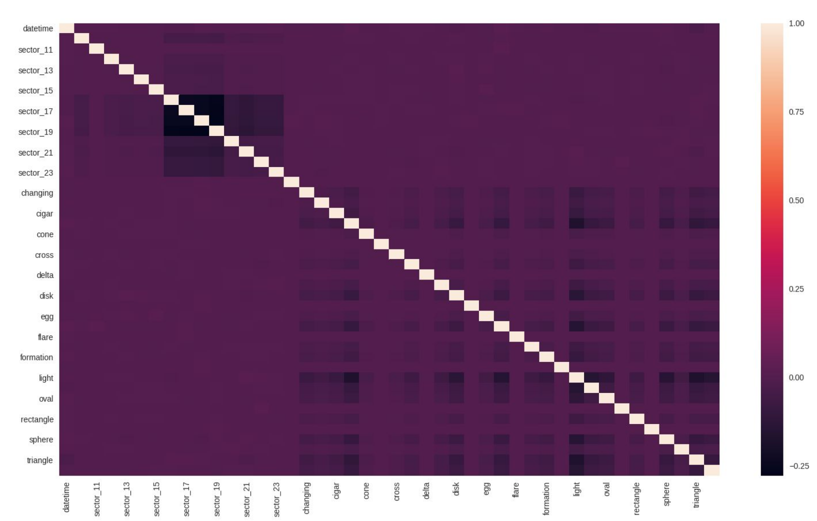
Sectors (incomplete)

count	7.839800e+04
mean	8.083026e+03
std	5.969395e+05
min	1.000000e-03
25%	3.000000e+01
50%	1.800000e+02
75%	6.000000e+02
max	9.783600e+07

Duration (seconds)

Note: Only a sample is included of the sectors and shapes

Table 2: Correlations



Correlation Matrix

Table 3: Full model results

```
==============================================================================
Dep. Variable:     duration (seconds)   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.1125
Date:                Sat, 09 Dec 2017   Prob (F-statistic):               1.00
Time:                        16:09:50   Log-Likelihood:            -1.1539e+06
No. Observations:               78398   AIC:                         2.308e+06
Df Residuals:                   78355   BIC:                         2.308e+06
Df Model:                          42
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
datetime    2075.0147   1075.473      1.929      0.054     -32.906    4182.935
sector_4     628.3120   4.56e+04      0.014      0.989   -8.88e+04       9e+04
sector_11     -0.1564   5.84e+05  -2.68e-07      1.000   -1.15e+06    1.15e+06
sector_12    -15.2953   5.54e+04     -0.000      1.000   -1.09e+05    1.08e+05
sector_13    -16.3449   4.84e+04     -0.000      1.000    -9.5e+04    9.49e+04
sector_14     -2.9525   5.31e+04  -5.56e-05      1.000   -1.04e+05    1.04e+05
sector_15     62.2251   5.22e+04      0.001      0.999   -1.02e+05    1.02e+05
sector_16     17.8981   3.76e+04      0.000      1.000   -7.36e+04    7.36e+04
sector_17   -454.5807   3.75e+04     -0.012      0.990    -7.4e+04    7.31e+04
sector_18   -470.5518   3.75e+04     -0.013      0.990    -7.4e+04     7.3e+04
sector_19   -621.7250   3.75e+04     -0.017      0.987   -7.41e+04    7.29e+04
sector_20   -203.7846   3.88e+04     -0.005      0.996   -7.62e+04    7.58e+04
sector_21    739.3642   3.84e+04      0.019      0.985   -7.44e+04    7.59e+04
sector_22    130.5959   3.89e+04      0.003      0.997   -7.62e+04    7.64e+04
sector_23   1696.8697    3.9e+04      0.044      0.965   -7.47e+04    7.81e+04
changed        0.0519   5.84e+05   8.89e-08      1.000   -1.15e+06    1.15e+06
changing     -99.7940   3.95e+04     -0.003      0.998   -7.75e+04    7.73e+04
chevron      -68.4953   4.17e+04     -0.002      0.999   -8.19e+04    8.18e+04
cigar       -108.5168   3.94e+04     -0.003      0.998   -7.73e+04    7.71e+04
circle      -157.3248   3.78e+04     -0.004      0.997   -7.43e+04     7.4e+04
cone         307.5768   4.96e+04      0.006      0.995   -9.69e+04    9.75e+04
crescent       0.3353   4.14e+05    8.1e-07      1.000   -8.11e+05    8.11e+05
cross        -17.8142   5.33e+04     -0.000      1.000   -1.05e+05    1.05e+05
cylinder     -47.5913   4.06e+04     -0.001      0.999   -7.97e+04    7.96e+04
delta         -0.4554   2.23e+05  -2.04e-06      1.000   -4.38e+05    4.38e+05
diamond      -71.0936   4.09e+04     -0.002      0.999   -8.03e+04    8.01e+04
disk        -290.9277   3.81e+04     -0.008      0.994   -7.49e+04    7.43e+04
dome          -0.0998   5.84e+05  -1.71e-07      1.000   -1.15e+06    1.15e+06
egg          -39.7947   4.28e+04     -0.001      0.999   -8.39e+04    8.39e+04
fireball    -189.1037   3.79e+04     -0.005      0.996   -7.45e+04    7.41e+04
flare         -0.1526   5.84e+05  -2.61e-07      1.000   -1.15e+06    1.15e+06
flash        -15.4910   4.05e+04     -0.000      1.000   -7.94e+04    7.94e+04
formation   -160.5014    3.9e+04     -0.004      0.997   -7.67e+04    7.64e+04
hexagon       -0.0174   5.84e+05  -2.98e-08      1.000   -1.15e+06    1.15e+06
light       1196.1690   3.75e+04      0.032      0.975   -7.23e+04    7.47e+04
other        974.2602    3.8e+04      0.026      0.980   -7.35e+04    7.55e+04
oval        -118.5368   3.84e+04     -0.003      0.998   -7.54e+04    7.52e+04
pyramid       -0.1325   5.84e+05  -2.27e-07      1.000   -1.15e+06    1.15e+06
rectangle    -92.0349   4.06e+04     -0.002      0.998   -7.96e+04    7.95e+04
round         -0.1843   4.14e+05  -4.45e-07      1.000   -8.11e+05    8.11e+05
sphere       987.2911   3.81e+04      0.026      0.979   -7.36e+04    7.56e+04
teardrop     -50.1808   4.29e+04     -0.001      0.999   -8.41e+04     8.4e+04
triangle    -401.5227   3.78e+04     -0.011      0.992   -7.45e+04    7.37e+04
unknown      -46.0445    3.8e+04     -0.001      0.999   -7.46e+04    7.45e+04
==============================================================================
Omnibus:                   334696.257   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):     993784937195.923
```

Full model results

Table 4: Reduced model results

```
==============================================================================
Dep. Variable:     duration (seconds)   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     15.16
Date:                Sat, 09 Dec 2017   Prob (F-statistic):           9.89e-05
Time:                        16:19:23   Log-Likelihood:            -1.1539e+06
No. Observations:               78398   AIC:                         2.308e+06
Df Residuals:                   78397   BIC:                         2.308e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
datetime    2085.7843    572.084      3.646      0.000     964.502    3207.066
==============================================================================
Omnibus:                   334696.541   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):     993789405497.758
```

Reduced model results