HI-SEAS Mars Solar Radiation Project
Terrance Randolph
IST 707 Data Mining

Introduction

“To Boldly Go Where No Man Has Gone Before…” (Star Trek) The mission to Mars for NASA astronauts is one of excitement yet extreme trepidations due to the long space journey and the tumultuous climates on Mars. Therefore, preparing for such a journey is paramount however, HI-SEAS (NASA Research Group) Mission IV & V were up for the challenge. The group of six individuals whose expertise ranges from Architecture, Physics, Astrobiology and Medicine spent twelve months in a Mars simulated environment. As a result, they were unable to leave their dome, slept in close quarters and had to ration food and water for eight to twelve months on the Mauna Loa Volcano side of Big Hawaii. This group of scientists spent their days collecting and understanding simulated cohabitation data from themselves as well as data from the actual Mars rovers, whereby, we are able to receive information of Mars solar radiation on any given day for four consecutive months.

The information concerning Mars solar radiation will assist in helping NASA make sound decisions. Hence, choices when it comes to building solar panels, dome habitats and space suits that are light, adaptable and sensible. Especially, when it comes to the weight and fuel for the long journey to Mars.

Analysis

Our first step was to arrange and clean up the data into a usable format. This took some time due to the size and scope of the data file and the format in which some of the data was logged. We were able to make our data set much more usable by removing N/A, grouping numeric values (such as Cardinal Directions) in to categories and making sure column names had uniform and more understandable headings Once that was completed we ran a regression analysis to identify which variables are significant in helping us to predict solar radiation levels. Using the significant variables identified, we applied multiple data mining tools including associative analysis, clustering and x to create a model to predict future solar radiation models.

The most challenging component of cleaning up our data was reducing the sample size of our data. The raw dataset had data recorded ever five minutes, 24 hours a day for 90 days. This gave us over 30,000 samples which very quickly proved to be a hassle. We needed to find a way to average out the data at some level to reduce the size of our data set without compromising the accuracy of the date. When we averaged out the data to hours and reran the regression to the data to be virtually the same as before it was averaged.

Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

Subsection 1: The Data

The dataset of Mars solar radiation from the months between September 2016 to December 2016 and is a viable analysis and prediction of what solar radiation impact would be given other factors and how astronauts should prepare. The data provide useful variables such as humidity, temperature, pressure and wind direction that we are able to analyze to determine which factor or combination of factors directly impact radiation if at all. Our goal with this dataset is to build a model to predict solar radiation levels on Mars down to the hour.

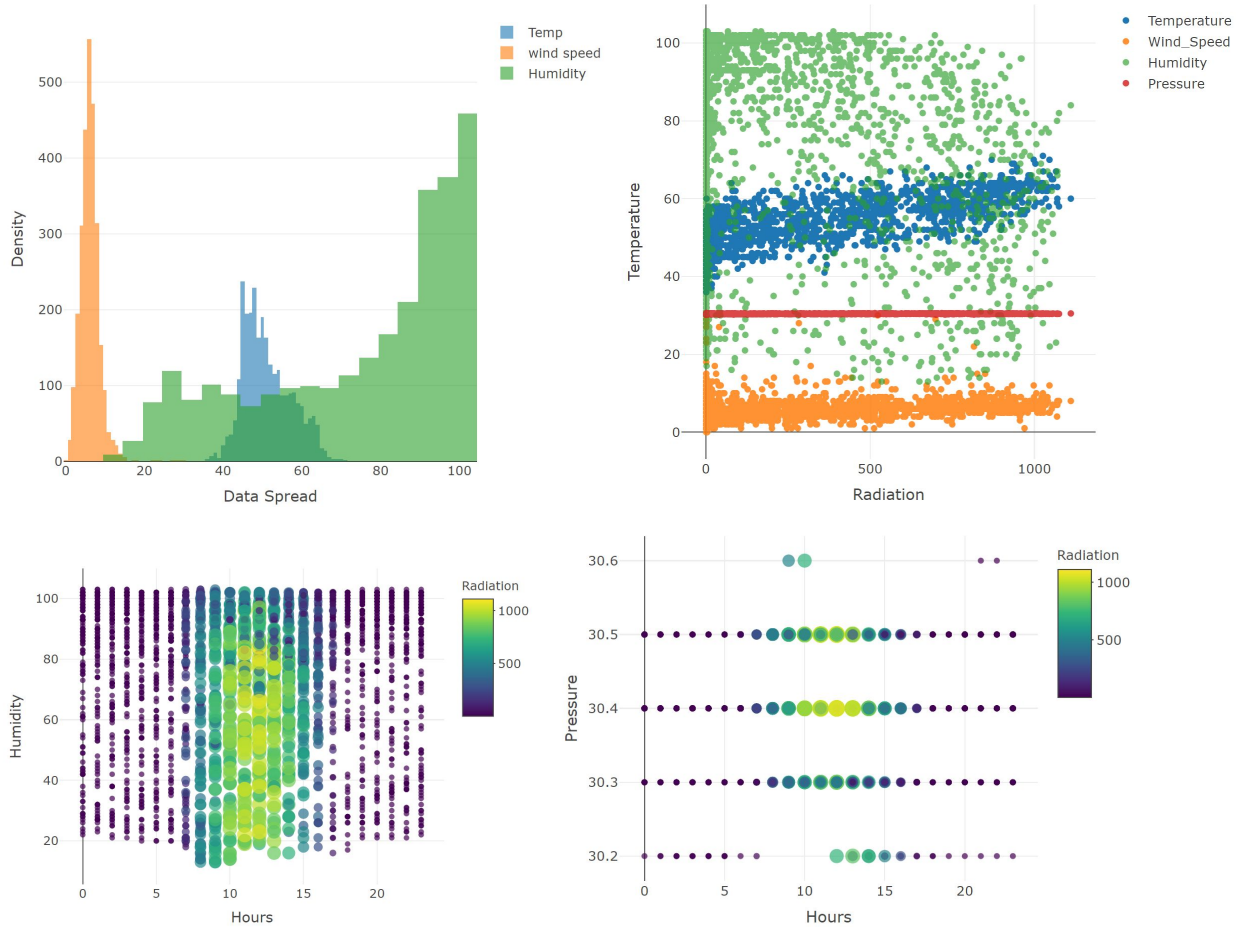
The data cleaning process for the solar radiation data set consisted of reading the (.csv) file into a data-frame then inspecting visually. I checked whether the variable names made sense, data types were correct and if columns needed to be removed. Shortly after the visual inspection it was time to clean the data by removing na's, splitting time date column into two columns, removing Unix variable and renaming the cleaned data-set. Although, the data seemed cleaned we wanted to aggregate the data by time, which as we soon found out would yield fruitful results. Therefore, the time column was recorded in 5 minute increments for all 24 hours of the day for over 90 days. Therefore, a new column was created from time data type to integer then, aggregated into hour increments resulting in 24 hours for each day, leaving almost 2,700 observations.

Next, the realization of creating two copies of the data set became apparent, one being the non-discrete data and the other being discrete data. The discreatized data consisted of most variables being parsed into 3 categories; low, medium and high. However, Radiation was broken into four levels due to the spread of the data. Most interestingly is the parsing of wind direction which was originally recorded as decimal degrees. Whereby, that unit of measurement needed to be translated into cardinal direction i.e. North & South. As a result, the decimal degrees were organized from 0 to 360, split into thirty-two and divided into seventeen for directions such as north-west or south-east. After conversion I realized that there were too many divisions therefore, the variable was aggregated into North, South, East and West.

Subsection 2: EDA

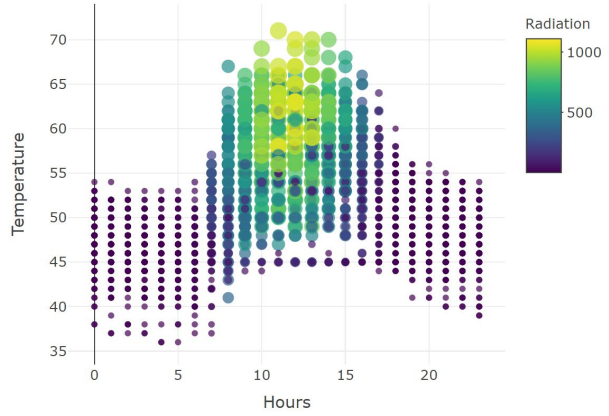
The exploration of the data was an integral portion of understanding the data shining a light on connection of most factors to Mars solar radiation levels. First gaining insights into the spread of the data gave us an understanding on much we could stretch our data, basically what are the limitations of the solar radiation daily. The analysis conveyed that temperature min/max difference is 35 with max/min being 71 - 36, pressure min/max difference is 0.4 with max/min being 30.6 - 30.2, Wind Speed min/max difference is 30 with max/min being 30 - 0, Wind

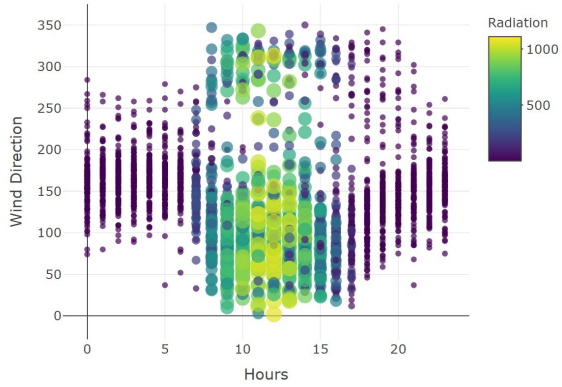
Direction min/max difference is 348 with max/min being 350 - 2 and Humidity min/max difference is 90 with max/min being 103 - 13. Next, with knowing the top and bottom of each variable in the data-set I moved on to visually displaying the variable to each other and individual variable compared to hours of the day. The graphs below told us a story that was later confirmed in our prediction models, in fact the plots themselves provide valuable answers to those solar radiation questions about Mars.

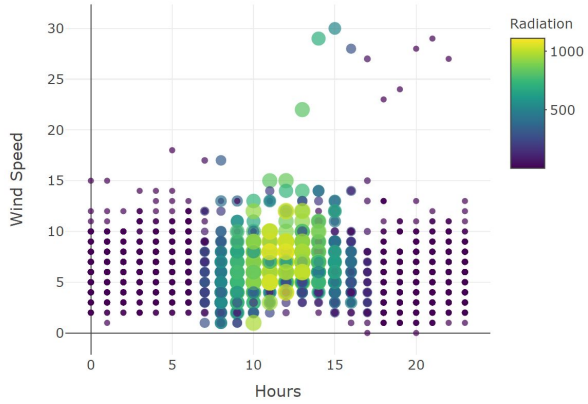


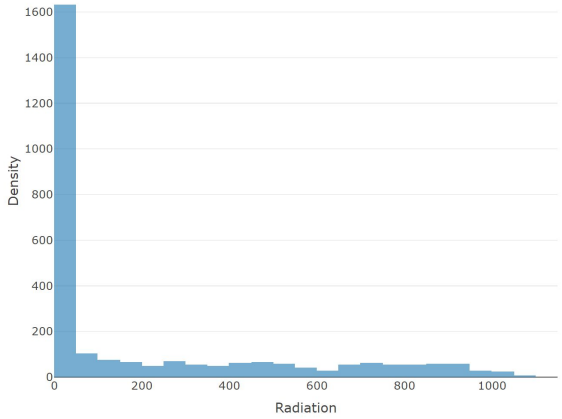
Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

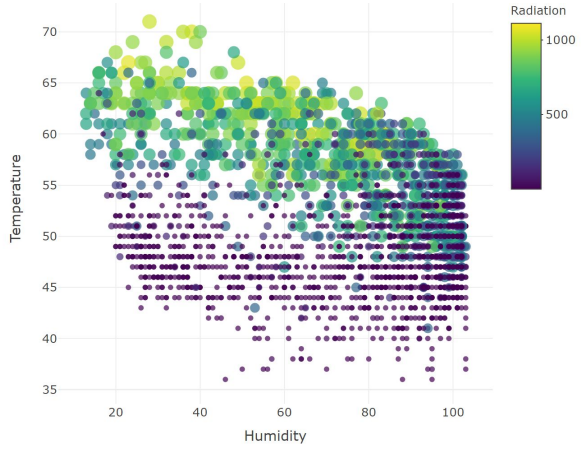
HI-SEAS Mars Solar Radiation Project
Terrance Randolph
IST 707 Data Mining

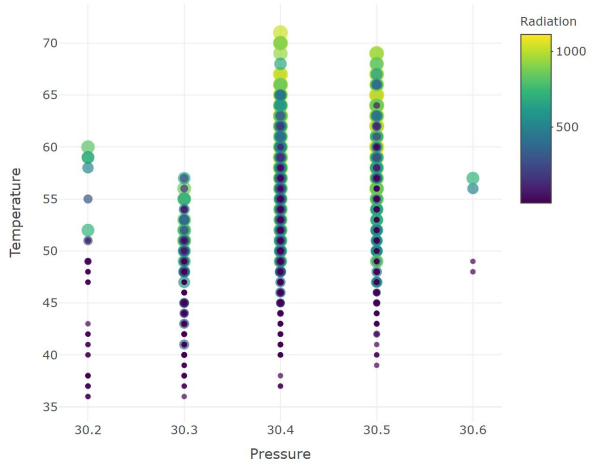




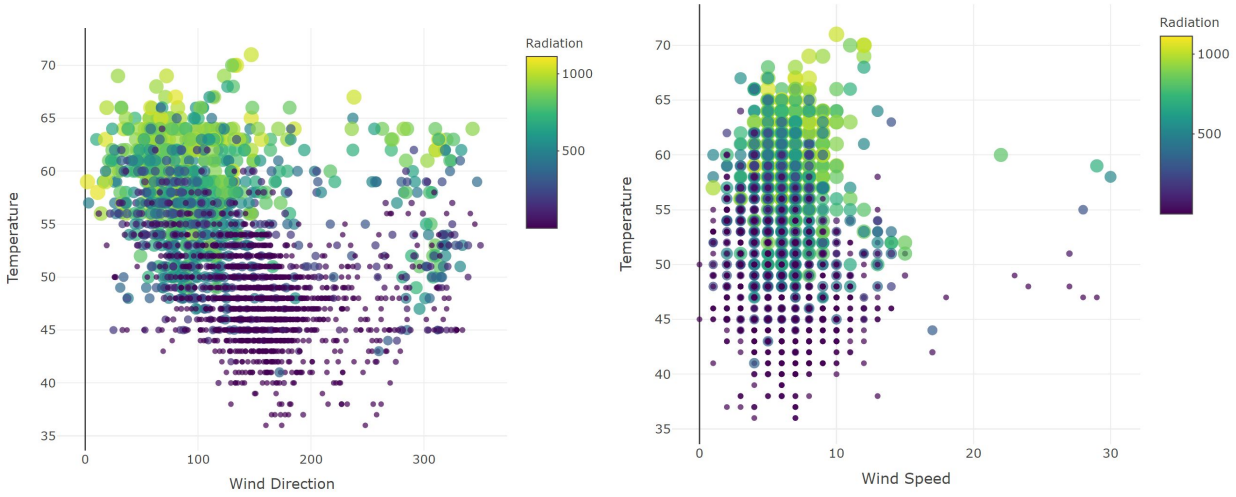








Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo



Subsections 3:

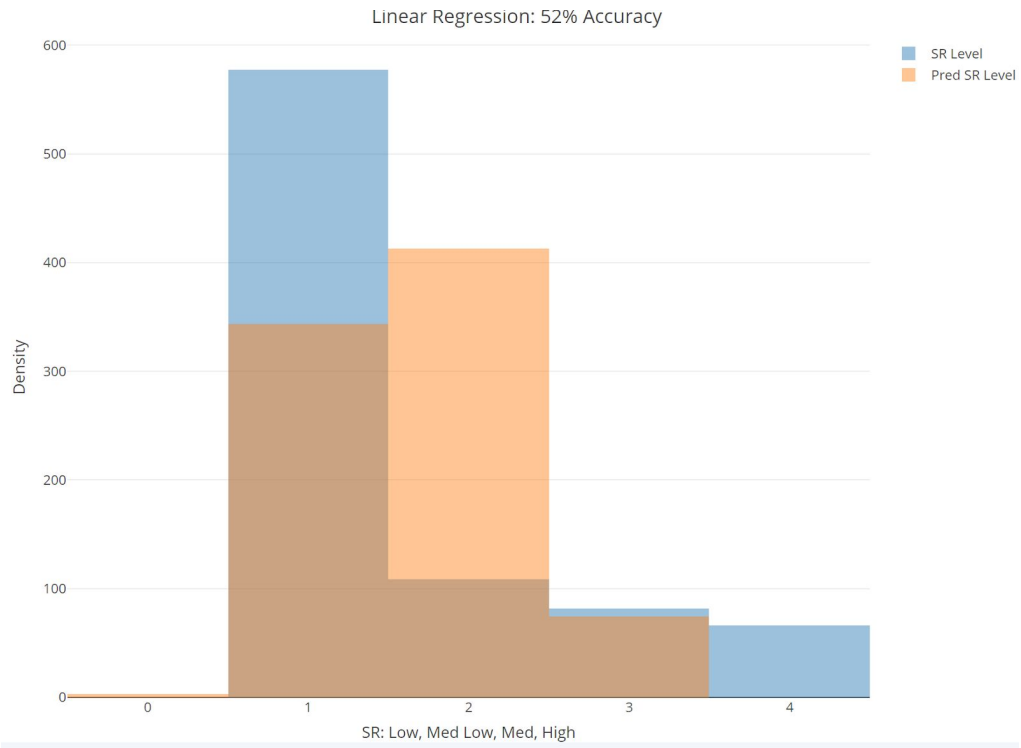
Regression

Before building a multi-regression model I need to inquire about the importance of each variable. Therefore, a Chi-Squared test was performed against each variable, concluding time of day(hours), humidity and temperature revealed themselves as being the most valuable factors. Moreover, the other columns possessed p-values below my alpha of 0.05, which convinced me that perhaps a regression model could be built with all variables included. Furthermore, after building the model and removing some higher p-value variables the accuracy of the regression model decreased from 53% to 45%. Therefore, it was injurious for the models to exclude those higher p-values.

Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

```
Call:
lm(formula = Radiation ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-748.77 -131.54  -19.83  102.92 1122.37

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.096e+04  8.325e+02  25.182  < 2e-16 ***
Temperature             3.817e+01  2.443e-01 156.266  < 2e-16 ***
Pressure               -7.460e+02  2.739e+01 -27.233  < 2e-16 ***
Humidity               -2.963e-01  5.754e-02  -5.149 2.64e-07 ***
Wind_Direction_Degrees -2.698e-01  1.737e-02 -15.534  < 2e-16 ***
Wind_Speed              7.901e+00  4.069e-01  19.418  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 207 on 22874 degrees of freedom
Multiple R-squared:  0.5679,    Adjusted R-squared:  0.5678
F-statistic:  6011 on 5 and 22874 DF,  p-value: < 2.2e-16
```

```
temperature Chi-squared against Radiation p-vlaue:     7.345324e-63
pressure Chi-squared against Radiation p-vlaue:        0.0384053
Humidity Chi-squared against Radiation p-vlaue:        8.245704e-50
Wind direction Chi-squared against Radiation p-vlaue:  2.384455e-318
wind speed Chi-squared against Radiation p-vlaue:      0.0002251407
time of day (hours) Chisquared against Radiation p-vlaue:    9.90174e-08
```

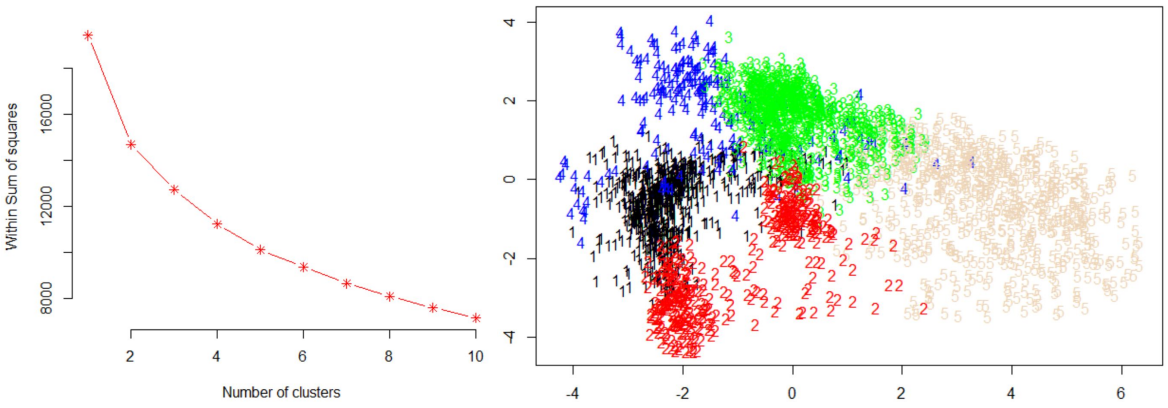


Clustering

Next was clustering analysis which was used to better understand what groups or relationships could be derived when trying to predict microgreys. In Addition, my primary purpose was to use this algorithm to show me visually if distinct clusters can be formed and where do they fall

Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

concerning microgrey levels. Whether, a logical cluster is formed along the different levels of microgreys (Mars Solar Radiation) or are the groups too intermingled for an accurate prediction. Luckily, after going from three kernels to fifteen, then finding comfort in five kernels due to kernel impact plot displaying shoulders (lines between cluster points) starting to shorten, informed my that five clusters are the optimal amount for good analysis.



	Radiation	Temperature	Pressure	Humidity	Wind_Direction_Degrees	Wind_Speed	Hours
1	-0.58690792	-0.7682009	-0.2991280	0.4807841	0.2565842	-0.11853023	-1.12184198
2	-0.57356043	-0.3653428	0.5087564	-1.5350673	0.3387613	0.57814563	-0.16779315
3	-0.39839058	-0.1060694	0.1218054	0.6233296	-0.3834019	-0.41745640	0.94152790
4	-0.09917262	-0.2880626	-1.4411899	0.4020136	2.1974598	0.76016382	0.42017008
5	1.59412719	1.3510601	0.3436155	-0.5258480	-0.7451994	0.05974488	0.02473516

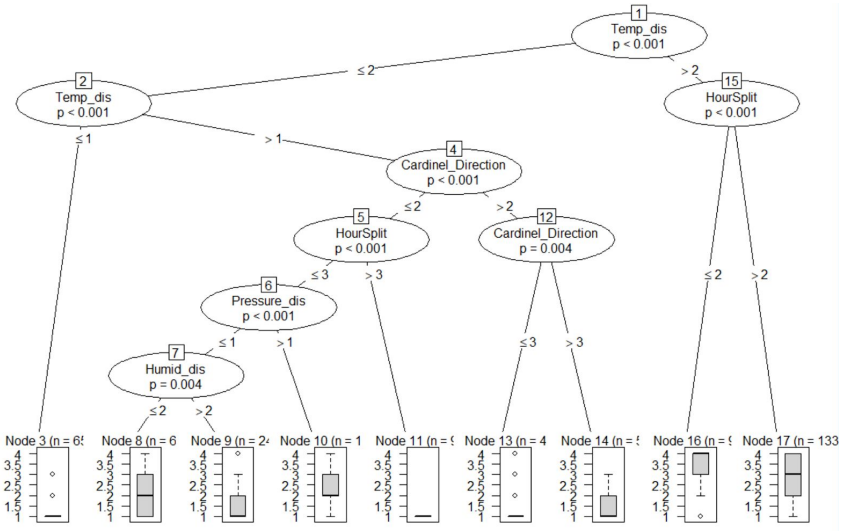
```
Within cluster sum of squares by cluster:
[1] 3529.266 5322.146 2643.738
 (between_SS / total_SS =  86.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

```
> clust_output$size
[1] 281 464 631 877 524
```

Decision Tree

The decision tree was one of the most shocking analyses, because the EDA has lead me along the path of time of day (hours) being the most important factor, However, the tree after being pruned a few times due to too many hours branches, the finds conveyed that temperature is a root node and if the temperature is low then humidity and wind direction decides how low radiation will be because, irrevocably if temperature is low radiation is almost always low. However, is temperature is medium or high then hours of sol (Martian day) is the deciding factor of how high radiation will be at that moment.

Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

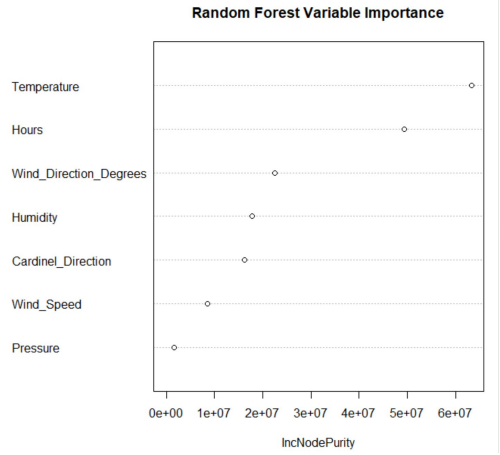


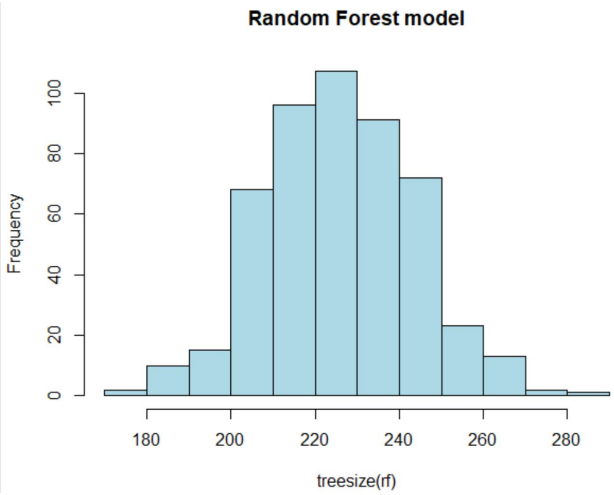
Random Forest

The discretized data-set was used to build the random forest model using libraries such as randomForest and RWeka. These libraries perform the same however, the libraries have additional functions that augment the analytical process which help the analyst gain a better understanding of the outcome. The quick and simple parameters from RWeka's random forest functions allowed me to easily set my trees and evaluate model strength based on cross validation. Moreover, randomForest library contain useful functions such as varImplot, which visualizes variable importance according to the y-variable/dependent variable (y-hat). Also, the getTree function grants information from a specific tree index. Therefore, this algorithm provided the most reasonable results with accuracy holding at 76% with 300 being the optimal number of trees.

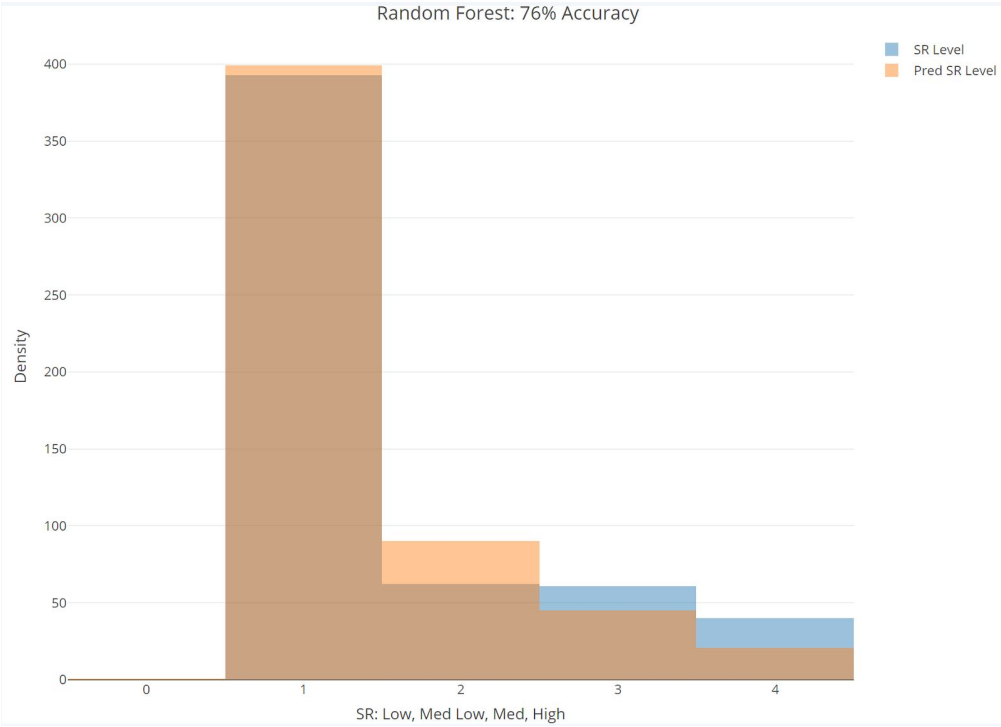
Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

HI-SEAS Mars Solar Radiation Project
Terrance Randolph
IST 707 Data Mining






```
Call:
 randomForest(formula = Radiation ~ ., data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 8034.579
                    % Var explained: 91.49
```



Support Vector Machine

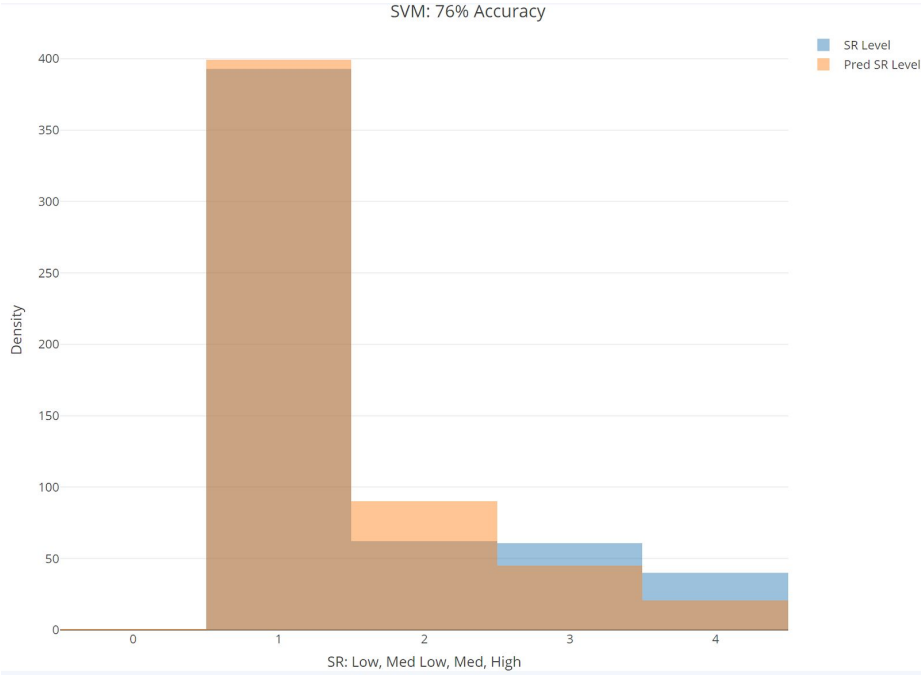
Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

HI-SEAS Mars Solar Radiation Project
Terrance Randolph
IST 707 Data Mining

When using support vector machine algorithm for predicting radiation levels on Mars and with the discretized data I built the model based off all variables; wind speed, hours, wind direction,temperature and humidity. However, i choose polynomial for the kernel because polynomials takes into account many terms/variables causing the model to be a bit malleable. Such flexibility rendered a 76% accuracy with cost maxing out around 500 with very little improvement shown with dramatic increases. Although, SVM rendered similar results as random forest, the additional parameter options and ways to improve the model sways in the favor of random forest being the best prediction method for Mars solar radiation predictions.

```
> SVM <- svm(RaidLevel~., data=train,
+                           kernel="polynomial", cost=1100,
+                           scale=FALSE)

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations
> ## COnfusion Matrix for training data to check model
> pred <- predict(SVM, test, type="class")
> (table(pred, test$RaidLevel))

pred      low low med med high
  low     551      68  49   11
  low med  16      15  13    4
  med       7       6  15    5
  high      4      10  16   34
```

```
> ###########polynomial with iris data small example
> ## Create a balanced Fed train
> SVM <- svm(RaidLevel~., data=train,
+                           kernel="polynomial", cost=2100,
+                           scale=FALSE)

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations

WARNING: reaching max number of iterations
> ## COnfusion Matrix for training data to check model
> pred <- predict(SVM, test, type="class")
> (table(pred, test$RaidLevel))

pred      low low med med high
  low     558      59  32   10
  low med   5      13  11    1
  med      12      18  33    8
  high      3       9  17   35
```



Interactive Dashboard Link:
https://datastudio.google.com/open/1a2FmLZnE--S_ul2l2Fq-STAldrp1DYYo

Results

The building of the training and testing set was the standard 70/30 split or 80/20 split for training and testing samples respectively. However, it was the consistent adjustment to parameters among other adjustments which included altering the variables levels for my discretized data-set. For example, I had to change the wind direction from factors of NE, SW and the other eight cardinal directions to four numeric directions with 1,2,3 and 4 being north, south, east and west. Next was the linear regression model whereby I was expecting the p-values to vary in significance, however all variable from the narrowed down data-set all had very high correlation to microgrey levels (Martian solar radiation levels). Forcing the regression model to be built with all five variables; wind speed, wind direction, pressure, humidity and hours.

Although, linear regression wasn't as ideal as expected the decision tree and clustering seemed to hold some prospects of running smoothly but, the first decision tree grew too many branches and needed pruning. Moreover, after pruning rpart did not want to comply with the discretize data therefore, printing the tree with a plot function displayed that the root is temperature which is followed by hours. Next, clustering muddled the initial three kernels but, slowly adding over fifteen kernels gave way to a more refined five kernels with distinguishable clusters. That have centers which explain how microgreys are associated with afternoon hours, humidity and high temperatures.

Lastly, the support vector machine and random forest were very similar in the end, not only in accuracy of 76% but cost and tree numbers. When building svm model with discrete data at 100 cost at polynomial kernels and results were mid 50% accuracy. However, after continuously changing the cost while keeping the malleable polynomial kernels ended in C = 500, because that is the threshold where accuracy growth is stagnant. In addition, random forest faced a similar problem of needing the number of trees to be 500 which coincide with the same accuracy result as svm. Moreover, figuring out a way of displaying accuracy has proven to be the most arduous of them all with predictions overlayed on test/actual solar radiation levels.

Conclusions:

Mars solar radiation is an interesting topic to cover that has yielded unexpected results and can prove to be valuable insights into Mars thin atmospheric problem and answer the question of how humans will survive on Mars. The questions of traveling to the moon was one doggedly pursued by many nations over 50 years ago with vast investments technologically and monetarily. However, at international collaboration was not as prevalent at that time because it

was a race for power and prestige. However, with that natural advancement in technology and common collaboration of minds from many companies and agencies across the globe, we as humans have fixed our gaze on Mars and its inhabitants. Therefore, understanding the impact Ionized radiation has on our non- iodized acclimated bodies due to the comforts of earth's thick atmosphere.

We received this data from NASA scientist and without understanding what information could be mined from the data, we proposed a simple yet practical question concerning how our analysis can be applied to real-world applications. The first question we proposed was concerning how much radiation Mars surface encounter on average on any given day or sol (Martian day) ? And the solar radiation levels averages 300 microgreys (Mars Radiation Unit of Measure). Such a large number can seriously harm humans without prolonged exposure. Therefore, is it possible for astronauts predict when the radiation will be unsafe? Yes, we have predicted with fair accuracy when microgreys will be at their highest during the day with data mining algorithms such as random forest. Such predictions allows NASA to properly test out the longevity of the rovers, solar panels, domes, martian suits etc.. and with our analysis especially graphs that clearly convey extremely high temperatures from 7 a.m to 6 p.m everyday. Therefore, the answer to our last question is yes, Mars solar radiation (micrograys) heavenly impact the mission, however, for only half of the day and any other out of dome activities can be pursued outside of those in order to limit high microgrey exposure.

In addition, the time constraints may not be easily overcome by waiting it out because, sunset and sunrise coincide with the high microgrey time of sol’s (Martian day’s). Whereby, NASA should improve environmental suits and rovers to withstand high microgreys. Lastly, the insights gained during this project may spawn many ideas and have many applications. For instance, mining asteroids is becoming a reality with U.S rovers and Japans hopping bots collecting minerals across our solar system the preparations for high radiation dovetailed with location and time can have some roots similar to our research. Therefore, humans understanding Martian microgreys could be the first step to sending humans “ to infinity and beyond…”. (Buzz Lightyear)

Citation

Interview with HI-SEAS & Ted Talk:
https://youtu.be/uZLvDi8uKDo
https://www.youtube.com/watch?v=uNYIAD601qY

NASA Martian Project:
https://www.nasa.gov/feature/goddard/real-martians-how-to-protect-astronauts-from-space-radiation-on-mars

HI-SEAS Mission & Home Page:
https://hi-seas.org/?page_id=6157
http://hi-seas.org/?page_id=5990

Radiation impact on Mars:
https://www.universetoday.com/14979/mars-radiation1/

Rover using RAD:
https://www.nasa.gov/jpl/msl/mars-rover-curiosity-pia17600.html#.XPFwfOkpA0M

Mars radiation:
Radiation Assessment Detector (RAD)
http://www.sci-news.com/space/science-mars-radiation-measurements-surface-01629.html

Astronaut prepare for mission to Mars
https://www.mars-one.com/faq/selection-and-preparation-of-the-astronauts/how-are-the-astronauts-prepared

Radiation Measurement:
https://www.nasa.gov/jpl/msl/mars-rover-curiosity-pia17600.html#.XP75pIhKjD4