Zipline project

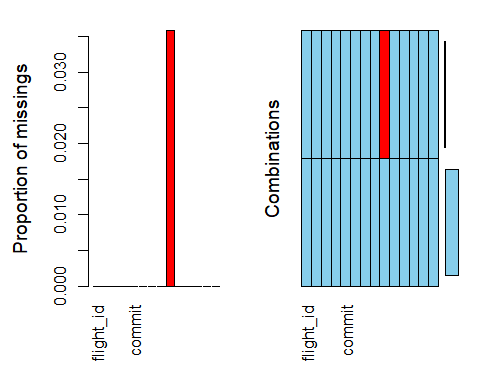
Pranav Sai Reddy Gudipally

May 4, 2019

## Data check for cleanliness

library(VIM)

aggr(mydata)

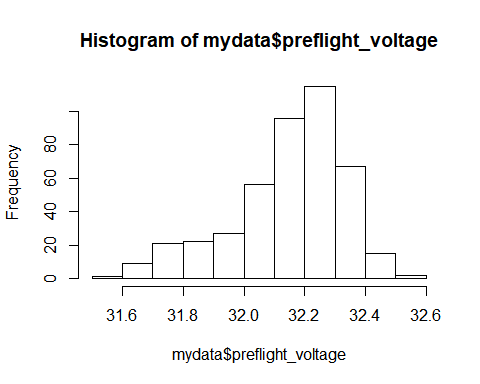


The data does not look clean as it seems to have some missing values, lets dig deeep and see how we can cleanse the data

## Data cleaning

#there seem to be missing values in one column of the dataset so we clean that variable(pre flight voltage) so we check the spread of that variable to see what we can replace the missing values with

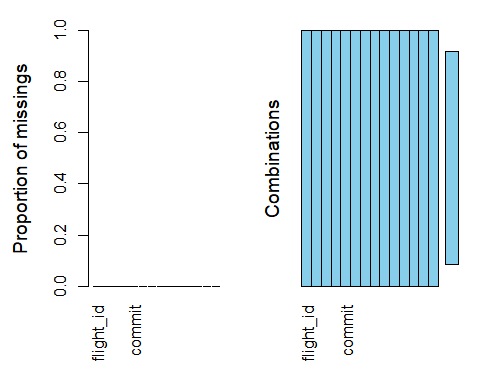
hist(mydata$preflight\_voltage) #the data in this column seems to be somewhat normal, so we replace the NA values in this column with the mean(average) which here is 32.14566493



mydata[is.na(mydata)] <- 32.14566493  
View(mydata)

## data check for cleanliness

aggr(mydata)

 The data seems to be completely clean now

# Exploratory Data Analysis

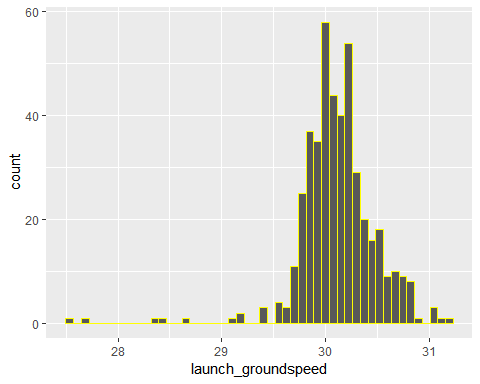
We now remove unwanted columns and non numeric data that we don’t need and plot the correlation matrix to see the correlation between diffferent variables to be able to perform effective analysis and plot visuals

reqdata<-mydata[-c(3,4,8,14)]#considering only required variables, naming the new data frame as "reqdata"  
View(reqdata)

# Visualizations

## Ground speed

library(ggplot2)  
ggplot(reqdata, aes(x=launch\_groundspeed))+  
 geom\_histogram(bins=50,color="yellow")

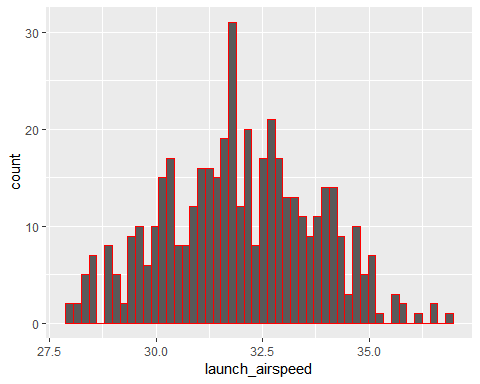


It is clear looking at the graph above that, ideal or average ground speed of the planes during launch is anywhere very very close to 30 meters/second

## 

## Air speed

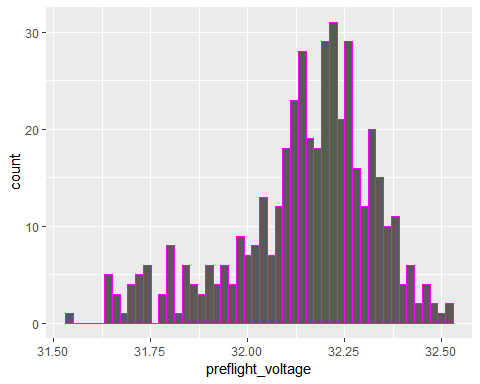
ggplot(reqdata, aes(x=launch\_airspeed))+  
 geom\_histogram(bins = 50,color="red")



Looking at the graph above, we can say that, considering that the flights are working at their best potential, their ideal or average air speed of the plane during launch is anywhere close to 32.5 meters/second

## Pre-flight Voltage

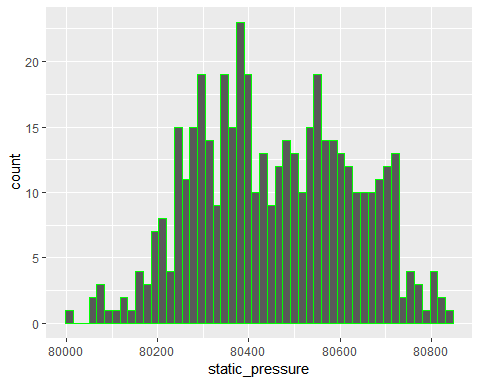
ggplot(reqdata, aes(x=preflight\_voltage))+  
 geom\_histogram(bins = 50,color="magenta")



We can see that dc voltage of the battery immediately prior to launch of most flights is around 32.35 volts

## Static Pressure

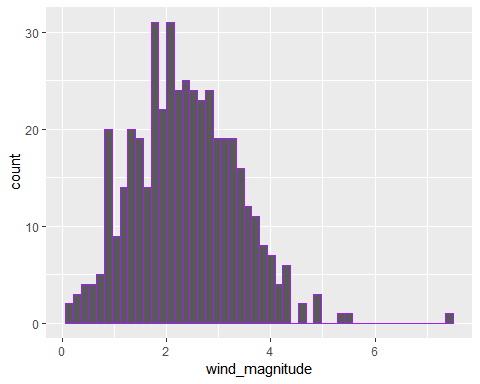
ggplot(reqdata, aes(x=static\_pressure))+  
 geom\_histogram(bins = 50,color="green")



We can look at the graph and interpret that the average static air pressure during launch for flights is around 80400 to 80600 pascals

## Wind magnitude

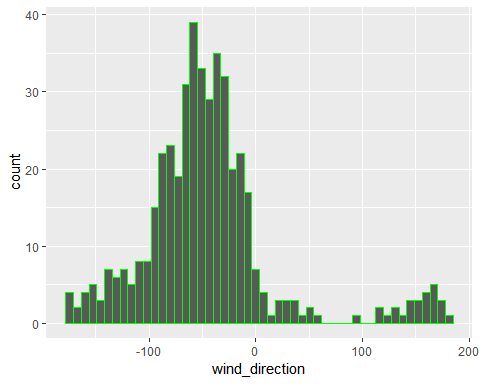
ggplot(reqdata, aes(x=wind\_magnitude))+  
 geom\_histogram(bins = 50,color="purple")



Average condition of magnitude of wing during launch is said to be around2 to 3 meters/second

## Wind Direction

ggplot(reqdata, aes(x=wind\_direction))+  
 geom\_histogram(bins = 50,color="green")



Wind direction on most of the occassions was to the North west direction during the take off or launch of our flights

### Looking at the above graphs we can infer that an average, normal or ideal condition for a flight is when

* Ground speed of the plane during launch is anywhere very very close to 30 meters/second
* Air speed of the plane during launch is anywhere close to 32.5 meters/second
* Direct Current voltage of the battery immediately prior to launch of most flights is around 32.35 volts
* Static air pressure during launch for flights is around 80400 to 80600 pascals
* magnitude of wing during launch is said to be around 2 to 3 meters/second
* Having wind direction North West

Correlation matrix generation for multi variable EDA

cordata<-reqdata[-c(3)]#removal of non numeric variables  
cor(cordata) #correlation matrix generation

## flight\_id air\_temperature launch\_airspeed  
## flight\_id 1.0000000000 -0.04017804 0.12414247  
## air\_temperature -0.0401780427 1.00000000 0.40686052  
## launch\_airspeed 0.1241424658 0.40686052 1.00000000  
## launch\_groundspeed 0.0001392636 -0.34804984 -0.38734258  
## preflight\_voltage 0.0531569571 0.08352864 0.06669974  
## rel\_humidity -0.2128251858 -0.60172131 -0.24507359  
## static\_pressure -0.1044500162 -0.25920114 0.11886280  
## wind\_direction -0.0097385400 -0.25374749 0.01325621  
## wind\_magnitude 0.1308604872 0.16762799 0.56072677  
## launch\_groundspeed preflight\_voltage rel\_humidity  
## flight\_id 0.0001392636 0.05315696 -0.21282519  
## air\_temperature -0.3480498367 0.08352864 -0.60172131  
## launch\_airspeed -0.3873425786 0.06669974 -0.24507359  
## launch\_groundspeed 1.0000000000 -0.04000662 0.07677876  
## preflight\_voltage -0.0400066242 1.00000000 -0.15723581  
## rel\_humidity 0.0767787561 -0.15723581 1.00000000  
## static\_pressure -0.1050345171 -0.15294884 0.54187845  
## wind\_direction 0.0368952562 -0.06957120 0.27452249  
## wind\_magnitude -0.1674123645 0.06121157 -0.26261923  
## static\_pressure wind\_direction wind\_magnitude  
## flight\_id -0.10445002 -0.00973854 0.13086049  
## air\_temperature -0.25920114 -0.25374749 0.16762799  
## launch\_airspeed 0.11886280 0.01325621 0.56072677  
## launch\_groundspeed -0.10503452 0.03689526 -0.16741236  
## preflight\_voltage -0.15294884 -0.06957120 0.06121157  
## rel\_humidity 0.54187845 0.27452249 -0.26261923  
## static\_pressure 1.00000000 0.17255502 -0.07191474  
## wind\_direction 0.17255502 1.00000000 -0.11138238  
## wind\_magnitude -0.07191474 -0.11138238 1.00000000

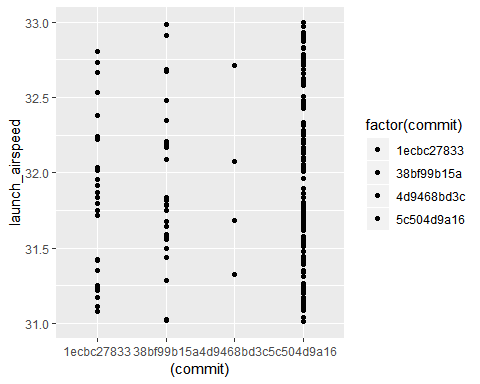
As we can see a good correlation between a few variables like wind magnitude and launch airspeed(56%), it would be better to find insights between them and the same can be done to other variables having a good enough correlation(or inverse correlation)

# Multi Variable Visualization

## Commit and Launch airspeed

ggplot(reqdata, aes(y=launch\_airspeed, x =(commit), fill= factor(commit))) +ylim(c(31,33))+  
 geom\_point()

## Warning: Removed 253 rows containing missing values (geom\_point).



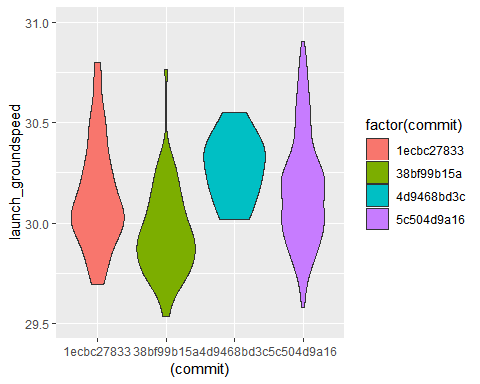
Looking at the plot above we can say that the sofware named “5c504d9a16” has been used the highest and is also the one giving good results, which is a launch airspeed value of 32.5 m/s

## 

## Commit and launch groundspeed

ggplot(reqdata, aes(y=launch\_groundspeed, x =(commit), fill= factor(commit))) +ylim(c(29.5,31))+  
 geom\_violin()

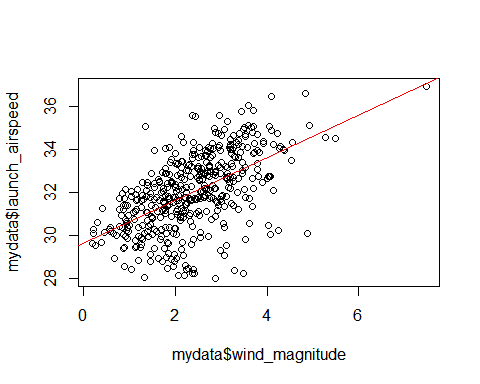
## Warning: Removed 16 rows containing non-finite values (stat\_ydensity).



Looking at the plot above we can say that the sofware named “1ecbc27833” is the one giving good results as it relates to average groundspeed of 30 m/s

## Launch air speed and wind magnitude

reg<-lm(mydata$launch\_airspeed~ mydata$wind\_magnitude)  
plot(mydata$wind\_magnitude , mydata$launch\_airspeed) + abline(reg,col="red")

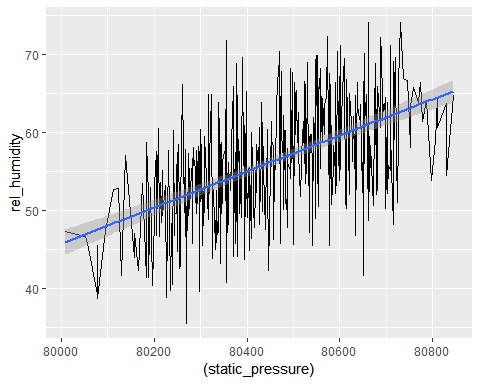


## integer(0)

Looking at the above plot it is clear that with increase in wind magnitude, there is an increase in launch air speed too, that means both of these variables are directly proportional to each other, and as most of the data point are around the line itself, there is no point to detect outliers here as it would do a negligible change to the line ultimately giving out similar insights

## Relative humidity and static pressure

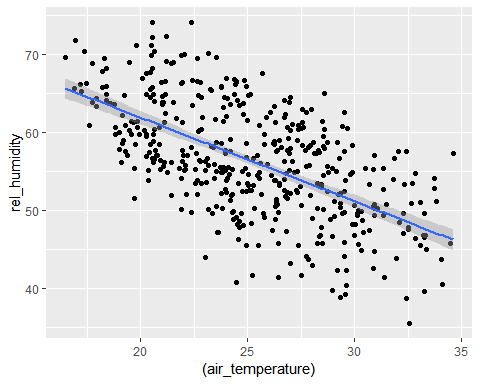
ggplot(reqdata, aes(y=rel\_humidity, x =(static\_pressure))) +  
 geom\_line()+geom\_smooth(method='lm')



We can see that, with increase in static pressure there is also an increase in relative humidity and vice versa, both of there variables are directly proportional to each other

## Relative humidity and air temperature

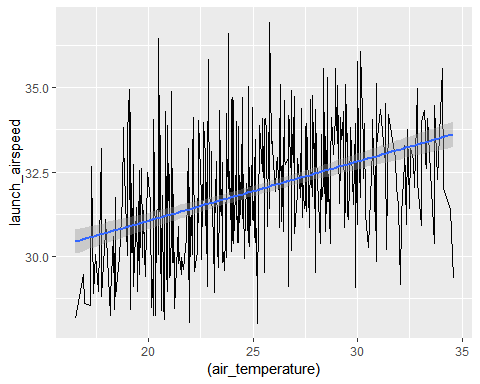
ggplot(reqdata, aes(y=rel\_humidity, x =(air\_temperature))) +  
 geom\_point()+geom\_smooth(method='lm')



We can see that, with decrease in relative humidity there is also a decrease in air temperature and vice versa, both of there variables are directly proportional to each other

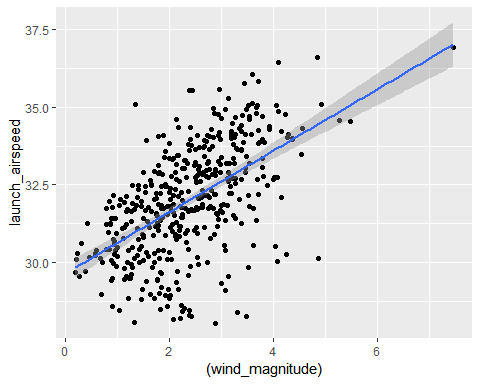
## Air temperature and air speed at launch

ggplot(reqdata, aes(y=launch\_airspeed, x =(air\_temperature))) +  
 geom\_line()+geom\_smooth(method='lm')



We can see that, with increase in temperature of the air there is also an increase in air speed of the plane during launch and vice versa, both of there variables are directly proportional to each other

ggplot(reqdata, aes(y=launch\_airspeed, x =(wind\_magnitude)))+  
 geom\_point()+geom\_smooth(method='lm')



We can see that, with increase in wind magnitude there is also an increase in air speed of plane during launch and vice versa, both of there variables are directly proportional to each other, outlier removal can be done for the above graph if needed by using the “xlim()” and “ylim()” functions but it would give the similar graph as we do not have extreme outlier density