DATA 606 Data Project Proposal

Kai Lukowiak

Data Preparation

```
library(dplyr)
library(tidyr)
library(data.table)
library(ggplot2)
library(ggthemes)
library(tibble)
library(knitr)
library(corrr)
```

```
test <- as.tibble(fread("/Users/kailukowiak/Data606_Proposal/test.csv", na.strings = c("-1","-1.0")))
```

```
##
Read 48.2% of 892816 rows
Read 88.5% of 892816 rows
Read 892816 rows and 58 (of 58) columns from 0.160 GB file in 00:00:04
```

```
train <- as.tibble(fread("/Users/kailukowiak/Data606_Proposal/train.csv", na.strings = c("-1","-1.0")))
head(test)
```

```
## # A tibble: 6 x 58
##      id ps_ind_01 ps_ind_02_cat ps_ind_03 ps_ind_04_cat ps_ind_05_cat
##   <int>     <int>         <int>     <int>         <int>         <int>
## 1     0         0             1         8             1             0
## 2     1         4             2         5             1             0
## 3     2         5             1         3             0             0
## 4     3         0             1         6             0             0
## 5     4         5             1         7             0             0
## 6     5         0             1         6             0             0
## # ... with 52 more variables: ps_ind_06_bin <int>, ps_ind_07_bin <int>,
## #   ps_ind_08_bin <int>, ps_ind_09_bin <int>, ps_ind_10_bin <int>,
## #   ps_ind_11_bin <int>, ps_ind_12_bin <int>, ps_ind_13_bin <int>,
## #   ps_ind_14 <int>, ps_ind_15 <int>, ps_ind_16_bin <int>,
## #   ps_ind_17_bin <int>, ps_ind_18_bin <int>, ps_reg_01 <dbl>,
## #   ps_reg_02 <dbl>, ps_reg_03 <dbl>, ps_car_01_cat <int>,
## #   ps_car_02_cat <int>, ps_car_03_cat <int>, ps_car_04_cat <int>,
## #   ps_car_05_cat <int>, ps_car_06_cat <int>, ps_car_07_cat <int>,
## #   ps_car_08_cat <int>, ps_car_09_cat <int>, ps_car_10_cat <int>,
## #   ps_car_11_cat <int>, ps_car_11 <int>, ps_car_12 <dbl>,
## #   ps_car_13 <dbl>, ps_car_14 <dbl>, ps_car_15 <dbl>, ps_calc_01 <dbl>,
## #   ps_calc_02 <dbl>, ps_calc_03 <dbl>, ps_calc_04 <int>,
## #   ps_calc_05 <int>, ps_calc_06 <int>, ps_calc_07 <int>,
## #   ps_calc_08 <int>, ps_calc_09 <int>, ps_calc_10 <int>,
## #   ps_calc_11 <int>, ps_calc_12 <int>, ps_calc_13 <int>,
## #   ps_calc_14 <int>, ps_calc_15_bin <int>, ps_calc_16_bin <int>,
## #   ps_calc_17_bin <int>, ps_calc_18_bin <int>, ps_calc_19_bin <int>,
## #   ps_calc_20_bin <int>
```

```
head(train)
```

```
## # A tibble: 6 x 59
##      id target ps_ind_01 ps_ind_02_cat ps_ind_03 ps_ind_04_cat
##   <int>  <int>     <int>         <int>     <int>         <int>
## 1     7      0         2             2         5             1
## 2     9      0         1             1         7             0
## 3    13      0         5             4         9             1
## 4    16      0         0             1         2             0
## 5    17      0         0             2         0             1
## 6    19      0         5             1         4             0
## # ... with 53 more variables: ps_ind_05_cat <int>, ps_ind_06_bin <int>,
## #   ps_ind_07_bin <int>, ps_ind_08_bin <int>, ps_ind_09_bin <int>,
## #   ps_ind_10_bin <int>, ps_ind_11_bin <int>, ps_ind_12_bin <int>,
## #   ps_ind_13_bin <int>, ps_ind_14 <int>, ps_ind_15 <int>,
## #   ps_ind_16_bin <int>, ps_ind_17_bin <int>, ps_ind_18_bin <int>,
## #   ps_reg_01 <dbl>, ps_reg_02 <dbl>, ps_reg_03 <dbl>,
## #   ps_car_01_cat <int>, ps_car_02_cat <int>, ps_car_03_cat <int>,
## #   ps_car_04_cat <int>, ps_car_05_cat <int>, ps_car_06_cat <int>,
## #   ps_car_07_cat <int>, ps_car_08_cat <int>, ps_car_09_cat <int>,
## #   ps_car_10_cat <int>, ps_car_11_cat <int>, ps_car_11 <int>,
## #   ps_car_12 <dbl>, ps_car_13 <dbl>, ps_car_14 <dbl>, ps_car_15 <dbl>,
## #   ps_calc_01 <dbl>, ps_calc_02 <dbl>, ps_calc_03 <dbl>,
## #   ps_calc_04 <int>, ps_calc_05 <int>, ps_calc_06 <int>,
## #   ps_calc_07 <int>, ps_calc_08 <int>, ps_calc_09 <int>,
## #   ps_calc_10 <int>, ps_calc_11 <int>, ps_calc_12 <int>,
## #   ps_calc_13 <int>, ps_calc_14 <int>, ps_calc_15_bin <int>,
## #   ps_calc_16_bin <int>, ps_calc_17_bin <int>, ps_calc_18_bin <int>,
## #   ps_calc_19_bin <int>, ps_calc_20_bin <int>
```

```
glimpse(train)
```

```
## Observations: 595,212
## Variables: 59
## $ id             <int> 7, 9, 13, 16, 17, 19, 20, 22, 26, 28, 34, 35, 3...
## $ target         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_01      <int> 2, 1, 5, 0, 0, 5, 2, 5, 5, 1, 5, 2, 2, 1, 5, 5,...
## $ ps_ind_02_cat  <int> 2, 1, 4, 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1,...
## $ ps_ind_03      <int> 5, 7, 9, 2, 0, 4, 3, 4, 3, 2, 2, 3, 1, 3, 11, 3...
## $ ps_ind_04_cat  <int> 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1,...
## $ ps_ind_05_cat  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_06_bin  <int> 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_07_bin  <int> 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1,...
## $ ps_ind_08_bin  <int> 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0,...
## $ ps_ind_09_bin  <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,...
## $ ps_ind_10_bin  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_11_bin  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_12_bin  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_13_bin  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_14      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_15      <int> 11, 3, 12, 8, 9, 6, 8, 13, 6, 4, 3, 9, 10, 12, ...
## $ ps_ind_16_bin  <int> 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0,...
## $ ps_ind_17_bin  <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_ind_18_bin  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1,...
```

```
## $ ps_reg_01      <dbl> 0.7, 0.8, 0.0, 0.9, 0.7, 0.9, 0.6, 0.7, 0.9, 0....
## $ ps_reg_02      <dbl> 0.2, 0.4, 0.0, 0.2, 0.6, 1.8, 0.1, 0.4, 0.7, 1....
## $ ps_reg_03      <dbl> 0.7180703, 0.7660777, NA, 0.5809475, 0.8407586,...
## $ ps_car_01_cat  <int> 10, 11, 7, 7, 11, 10, 6, 11, 10, 11, 11, 11, 6,...
## $ ps_car_02_cat  <int> 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1,...
## $ ps_car_03_cat  <int> NA, NA, NA, 0, NA, NA, NA, 0, NA, 0, NA, NA, NA...
## $ ps_car_04_cat  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 8, 0, 0, 0, 0, 9,...
## $ ps_car_05_cat  <int> 1, NA, NA, 1, NA, 0, 1, 0, 1, 0, NA, NA, NA, 1,...
## $ ps_car_06_cat  <int> 4, 11, 14, 11, 14, 14, 11, 11, 14, 14, 13, 11, ...
## $ ps_car_07_cat  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ ps_car_08_cat  <int> 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0,...
## $ ps_car_09_cat  <int> 0, 2, 2, 3, 2, 0, 0, 2, 0, 2, 2, 0, 2, 2, 2, 0,...
## $ ps_car_10_cat  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ ps_car_11_cat  <int> 12, 19, 60, 104, 82, 104, 99, 30, 68, 104, 20, ...
## $ ps_car_11      <int> 2, 3, 1, 1, 3, 2, 2, 3, 3, 2, 3, 3, 3, 3, 1, 2,...
## $ ps_car_12      <dbl> 0.4000000, 0.3162278, 0.3162278, 0.3741657, 0.3...
## $ ps_car_13      <dbl> 0.8836789, 0.6188165, 0.6415857, 0.5429488, 0.5...
## $ ps_car_14      <dbl> 0.3708099, 0.3887158, 0.3472751, 0.2949576, 0.3...
## $ ps_car_15      <dbl> 3.605551, 2.449490, 3.316625, 2.000000, 2.00000...
## $ ps_calc_01     <dbl> 0.6, 0.3, 0.5, 0.6, 0.4, 0.7, 0.2, 0.1, 0.9, 0....
## $ ps_calc_02     <dbl> 0.5, 0.1, 0.7, 0.9, 0.6, 0.8, 0.6, 0.5, 0.8, 0....
## $ ps_calc_03     <dbl> 0.2, 0.3, 0.1, 0.1, 0.0, 0.4, 0.5, 0.1, 0.6, 0....
## $ ps_calc_04     <int> 3, 2, 2, 2, 2, 3, 2, 1, 3, 2, 2, 2, 4, 2, 3, 2,...
## $ ps_calc_05     <int> 1, 1, 2, 4, 2, 1, 2, 2, 1, 2, 3, 2, 1, 1, 1, 1,...
## $ ps_calc_06     <int> 10, 9, 9, 7, 6, 8, 8, 7, 7, 8, 8, 8, 8, 10, 8, ...
## $ ps_calc_07     <int> 1, 5, 1, 1, 3, 2, 1, 1, 3, 2, 2, 2, 4, 1, 2, 5,...
## $ ps_calc_08     <int> 10, 8, 8, 8, 10, 11, 8, 6, 9, 9, 9, 10, 11, 8, ...
## $ ps_calc_09     <int> 1, 1, 2, 4, 2, 3, 3, 1, 4, 1, 4, 1, 1, 3, 3, 2,...
## $ ps_calc_10     <int> 5, 7, 7, 2, 12, 8, 10, 13, 11, 11, 7, 8, 9, 8, ...
## $ ps_calc_11     <int> 9, 3, 4, 2, 3, 4, 3, 7, 4, 3, 6, 9, 6, 2, 4, 5,...
## $ ps_calc_12     <int> 1, 1, 2, 2, 1, 2, 0, 1, 2, 5, 3, 2, 3, 0, 1, 2,...
## $ ps_calc_13     <int> 5, 1, 7, 4, 1, 0, 0, 3, 1, 0, 3, 1, 3, 4, 3, 6,...
## $ ps_calc_14     <int> 8, 9, 7, 9, 3, 9, 10, 6, 5, 6, 6, 10, 8, 3, 9, ...
## $ ps_calc_15_bin <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ ps_calc_16_bin <int> 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1,...
## $ ps_calc_17_bin <int> 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1,...
## $ ps_calc_18_bin <int> 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ ps_calc_19_bin <int> 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1,...
## $ ps_calc_20_bin <int> 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0,...
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Can simple statistical and ML models predict accidents better than baisic markers like age adgusted rates or the base line rate?

Cases

What are the cases, and how many are there?

The cases are individual people who bought insurance.

```
x = nrow(train) + nrow(test)
```

There are: 1488028 cases.

Data collection

Describe the method of data collection.

Data collection was easy since the data sets were posted on kaggle. Files can be downloaded here: https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data

Type of study

What type of study is this (observational/experiment)?

This is an observational study based on insurance claims and people's attributes.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data

Response

What is the response variable, and what type is it (numerical/categorical)?

The response variable in categorical because it takes on a value of zero and 1 if there was a claim.

Explanatory

What is the explanatory variable, and what type is it (numerical/categorival)?

There are a mix of numeric and categorical variables (nothing logistic regression / random forrests can't handle)

Relevant summary statistics

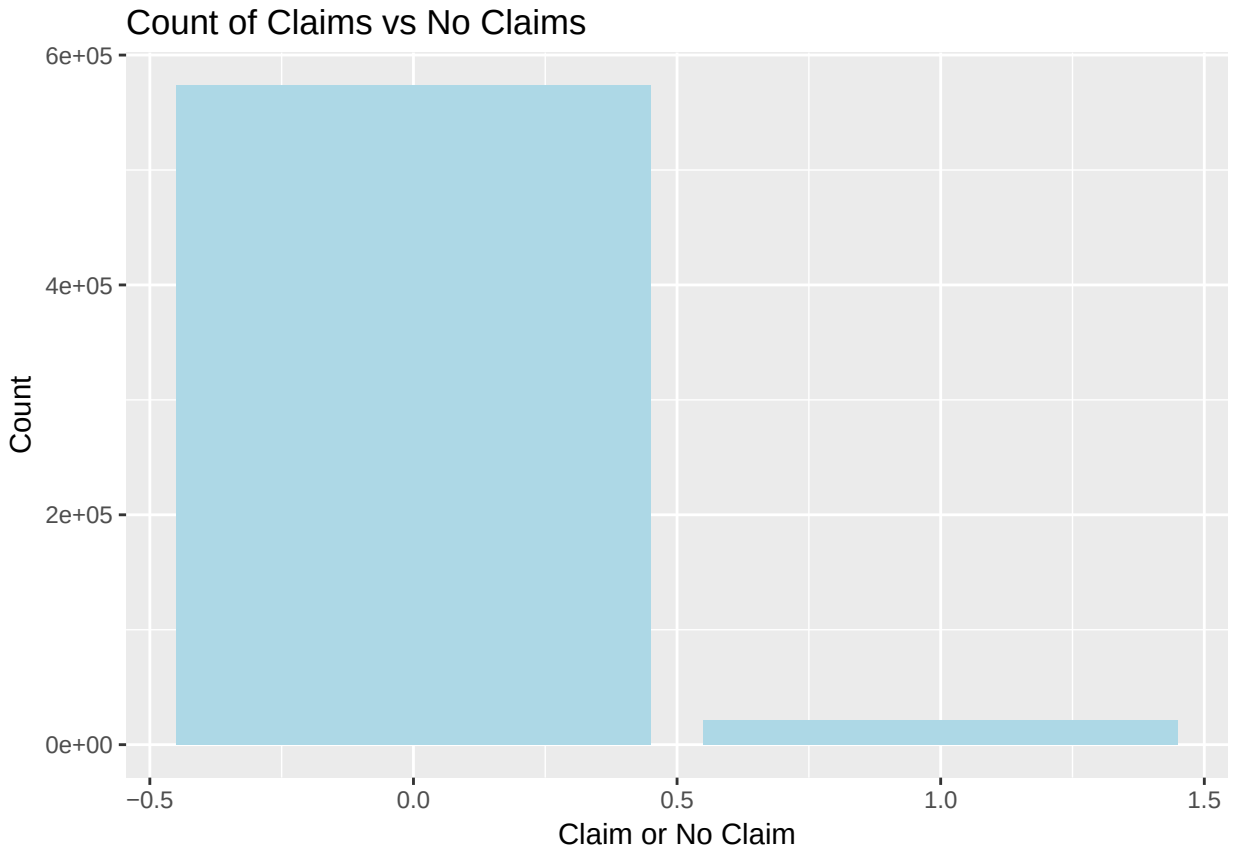
Provide summary statistics relevant to your research question. For example, if you're comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
y = 1
```

There are a lot of NAs.

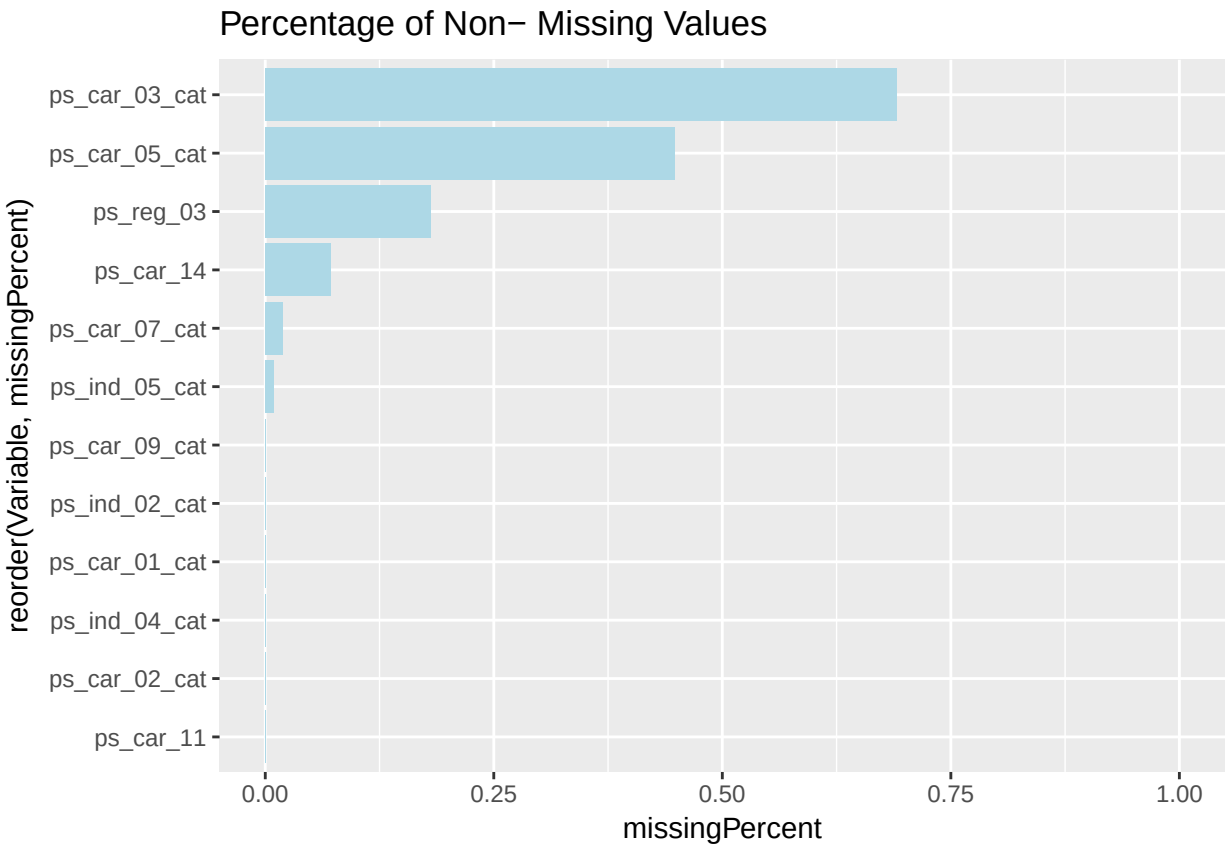
```
train %>%
  select(target) %>%
  group_by(target) %>%
  summarise(n = n()) %>%
  ggplot( aes(x = target, y = n))+
  geom_bar(stat = 'identity', fill = 'light blue') +
  ggtitle("Count of Claims vs No Claims") +
```

```
  xlab('Claim or No Claim') +
  ylab('Count')
```




```
naVals <- test %>%
  select(which(colMeans(is.na(.)) > 0)) %>%
  summarise_all(funs(sum(is.na(.))/n())) %>%
  gather(key = "Variable", value = "missingPercent")
```

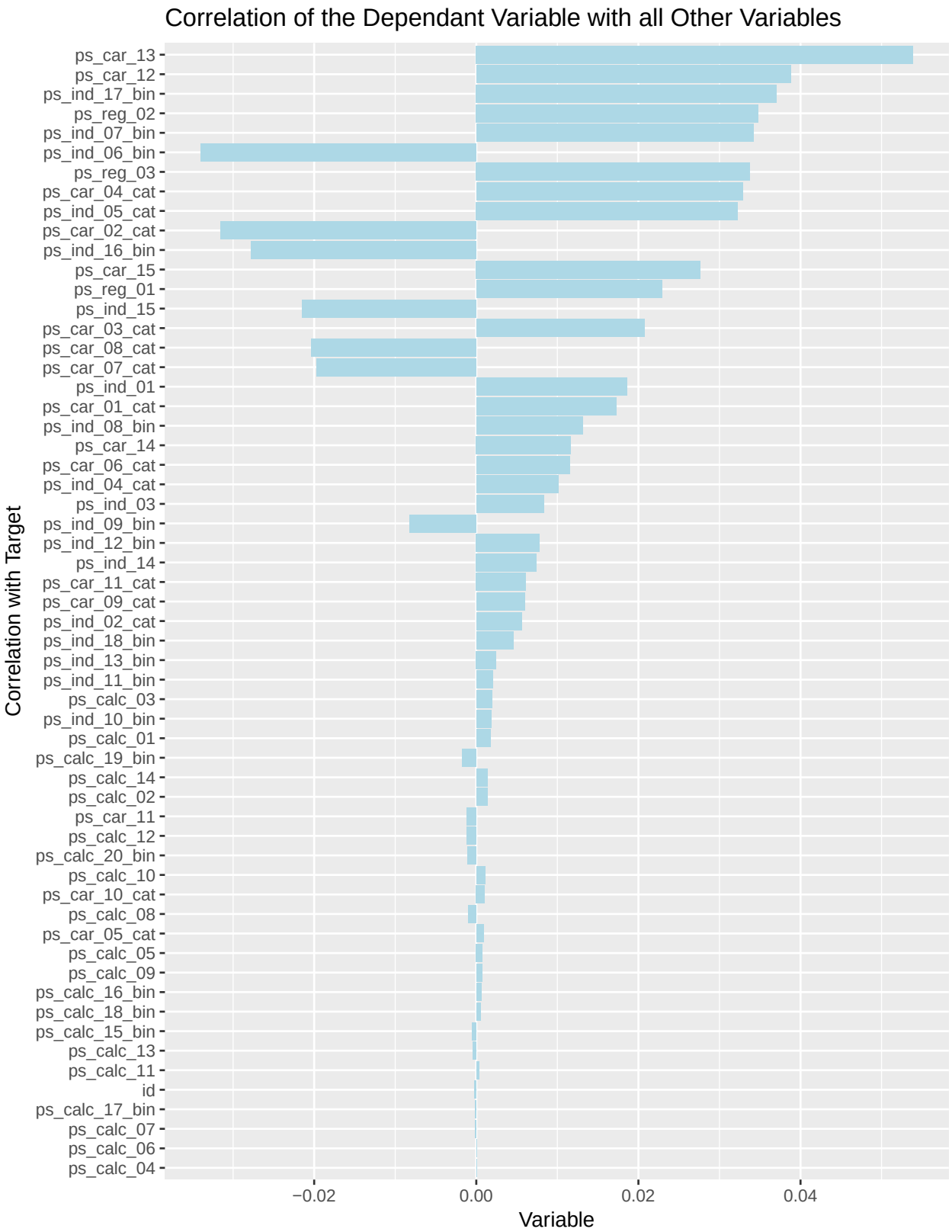
```
ggplot(naVals, aes(x = reorder( Variable, missingPercent), y = missingPercent)) +
  geom_bar(stat = "identity", fill = 'light blue') +
  ylim(0,1) +
  ggtitle("Percentage of Non- Missing Values") +
  coord_flip()
```



We can see that most of `ps_car_03_cat` and `ps_car_05_cat` is missing. Other than that, most variables contain few if any NAs.

```
corrDF <- train %>%
  correlate() %>%
  focus(target)
```

```
ggplot(corrDF, aes(x =reorder(rowname, abs(target)), y = target)) +
  geom_bar(stat = 'identity', fill = 'light blue') +
  coord_flip()+
  ylab('Variable')+
  xlab('Correlation with Target')+
  ggtitle('Correlation of the Dependant Variable with all Other Variables')
```



Because the correlation values are so low, we will probably have to equalize the success and failure rates so that success doesn't dominate the results.