FINAL EXAM

Mario Santana

March 15, 2016

When approaching the final, at first I thought it would be best to run regressions in order to build the relationship but the more I ran the regressions the more it seemed like the relationships where, for the most part, weak. They weak in the sense that, even though the coefficients were statistically significant, the coefficients were very small and did not impact the variable of interest. So I decided to take the approach of using bar graphs since most variables of interest were not ideal for regression.

Part a)

Relationship between Departure Delays and Weather

```
summary(reg_flights.weather)
```

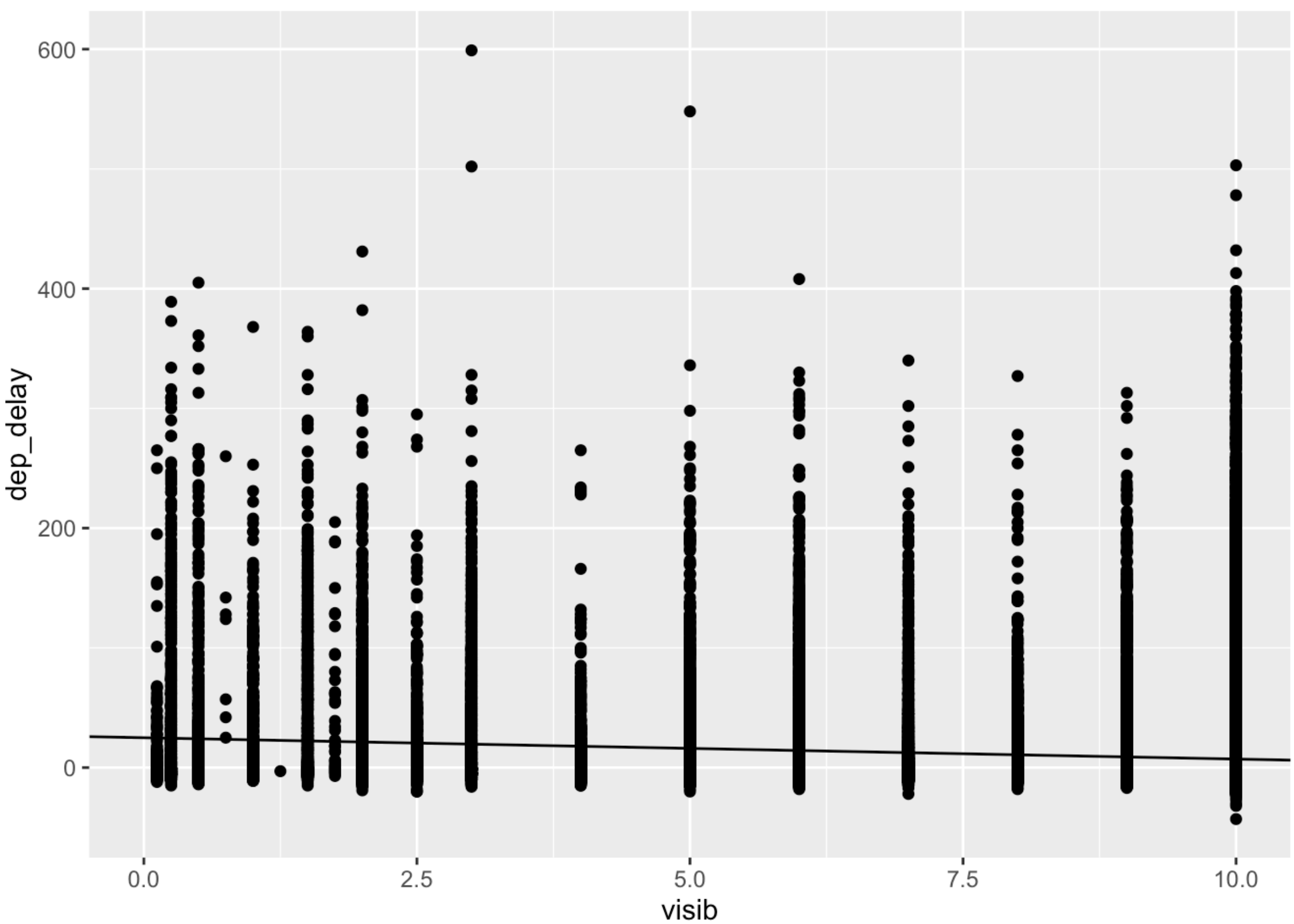
```
##
## Call:
## lm(formula = flights_weather$dep_delay ~ flights_weather$visib +
##     flights_weather$precip + flights_weather$pressure + flights_weather$wind_speed
+
##     flights_weather$wind_dir + flights_weather$humid + flights_weather$dewp +
##     flights_weather$temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -48.95  -12.85   -8.69   -1.42 1299.92
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                301.904134  17.682700  17.073  < 2e-16 ***
## flights_weather$visib       -1.815604   0.082013 -22.138  < 2e-16 ***
## flights_weather$precip     -37.153537   6.007747  -6.184 6.27e-10 ***
## flights_weather$pressure    -0.260790   0.016441 -15.862  < 2e-16 ***
## flights_weather$wind_speed   0.356078   0.023278  15.297  < 2e-16 ***
## flights_weather$wind_dir    -0.007568   0.001132  -6.687 2.29e-11 ***
## flights_weather$humid       -0.140730   0.037831  -3.720 0.000199 ***
## flights_weather$dewp         0.121564   0.080581   1.509 0.131407
## flights_weather$temp        -0.189169   0.074609  -2.535 0.011231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.74 on 86508 degrees of freedom
##   (13483 observations deleted due to missingness)
## Multiple R-squared:  0.01533,    Adjusted R-squared:  0.01524
## F-statistic: 168.4 on 8 and 86508 DF,  p-value: < 2.2e-16
```

From these results we can see that variables that seem to have an effect on departure delays are visibility and precipitation. All the other variables, even though they are statistically sinificant, dont effect departure delays much.

Regressing departure delays on visibility will give us a better understanding on the relationship between delays and visibility

```
summary(lm(dep_delay ~ visib, data = flights_weather))
```

```
##
## Call:
## lm(formula = dep_delay ~ visib, data = flights_weather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -50.09  -13.61  -10.09   -2.08 1293.91
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.97128    0.41820   59.71   <2e-16 ***
## visib       -1.78809    0.04508  -39.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.9 on 99998 degrees of freedom
## Multiple R-squared:  0.01549,    Adjusted R-squared:  0.01548
## F-statistic:  1573 on 1 and 99998 DF,  p-value: < 2.2e-16
```

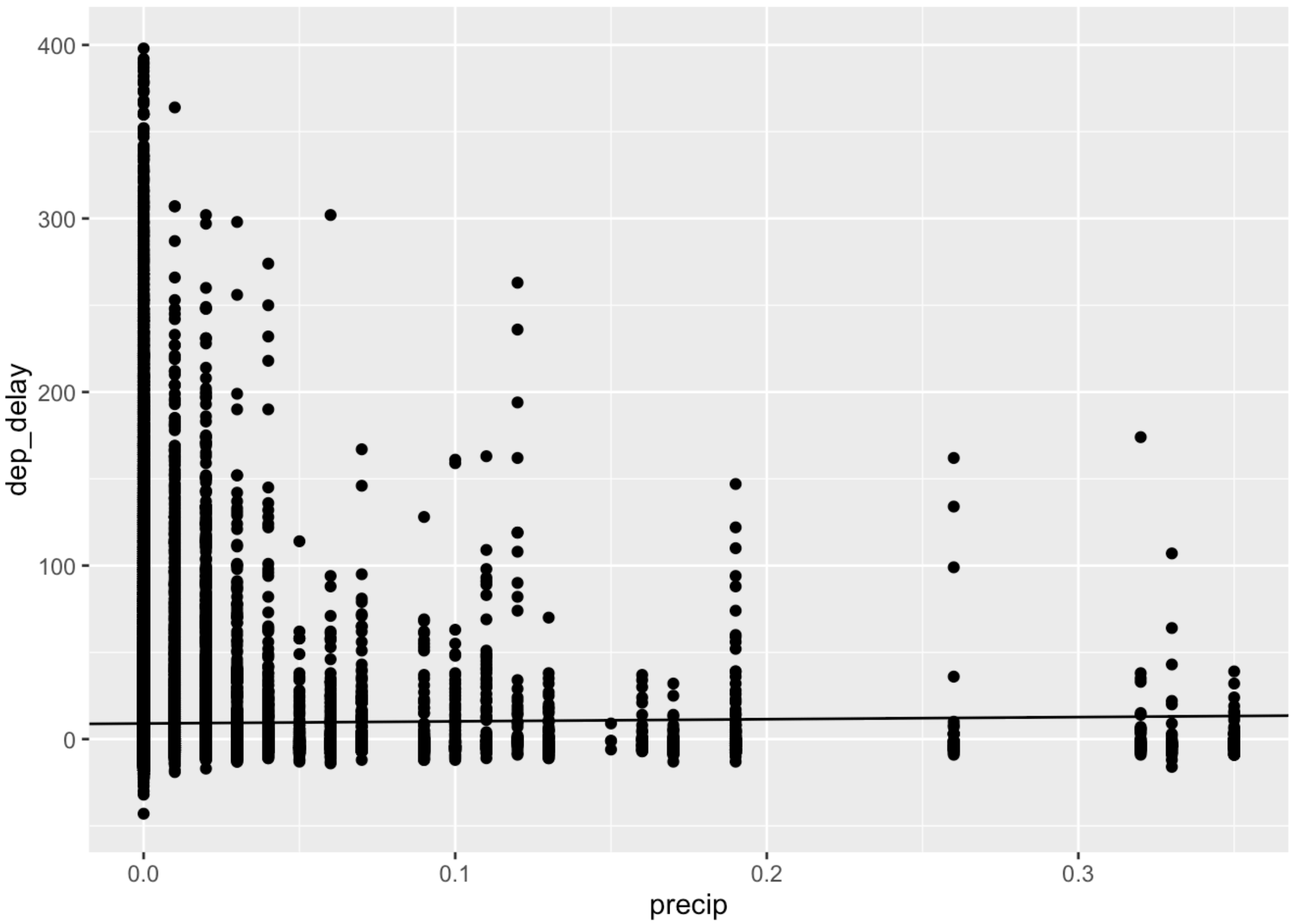


This shows that the relationship between visibility and departure delays is negative. It is statistically significant as well. Therefore, when visibility increases by one unit departure delay decreases by about 2 units.

Now Regressing departure delays on precipitation will give us a better understanding on the relationship between departure delays and precipitation

```
summary(lm(dep_delay ~ precip, data = flights_weather))
```

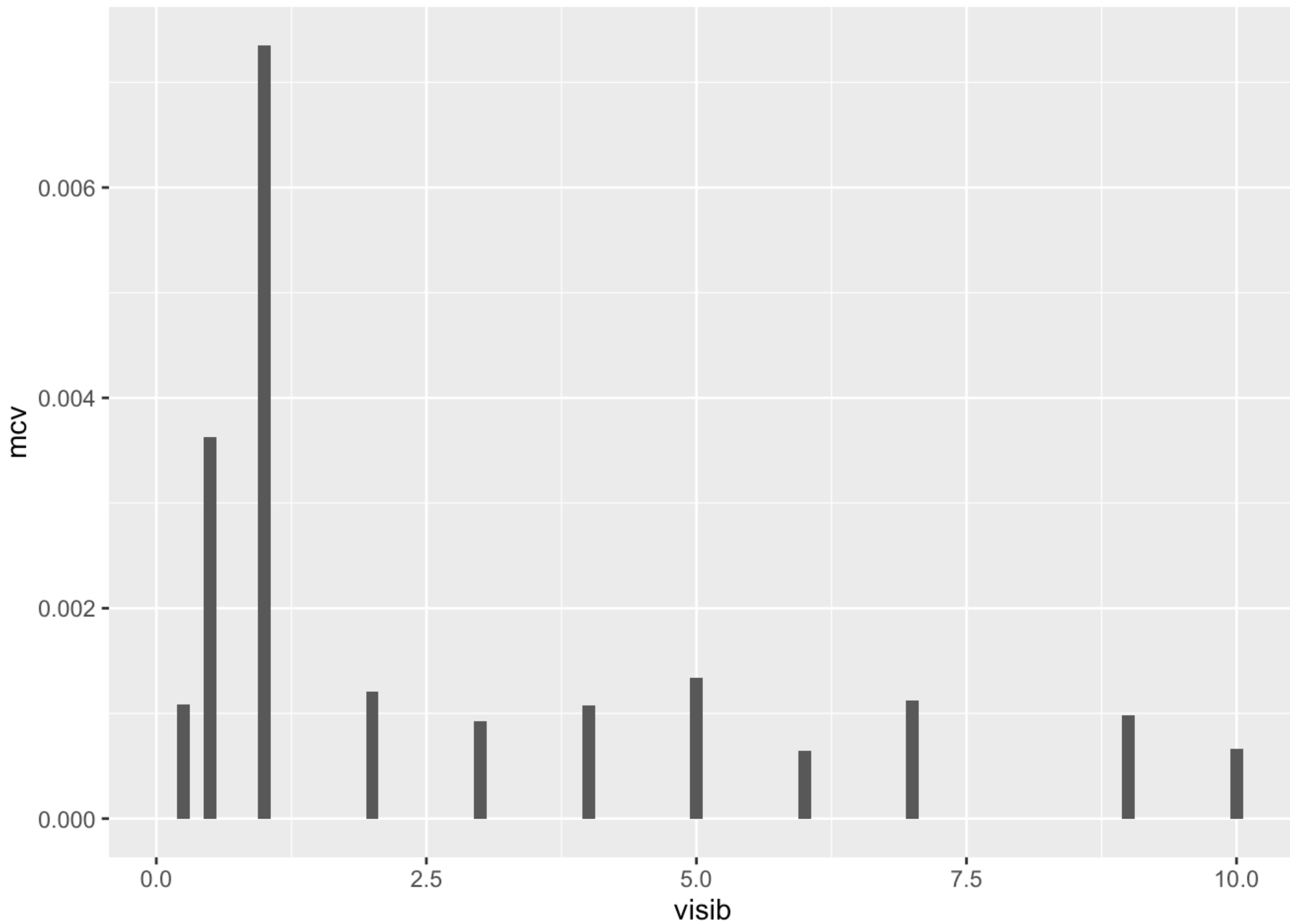
```
##
## Call:
## lm(formula = dep_delay ~ precip, data = flights_weather)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -51.91  -13.91  -10.91   -1.91 1292.09
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.9082     0.1089  81.797   <2e-16 ***
## precip       12.5060     5.8920   2.123   0.0338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.17 on 99998 degrees of freedom
## Multiple R-squared:  4.505e-05,  Adjusted R-squared:  3.505e-05
## F-statistic: 4.505 on 1 and 99998 DF,  p-value: 0.0338
```



This shows that the relationship between precipitation and departure delays is positive. It is also statistically significant. Therefore, when precipitation increases by one unit departure delay increases by about 12 units. Precipitation has a stronger effect on departure delay than visibility does.

Relationship between Cacelations and Weather

```
ggplot(mean_canceled_visibility, aes(visib, y = mcv)) + geom_bar(stat = "identity")
```



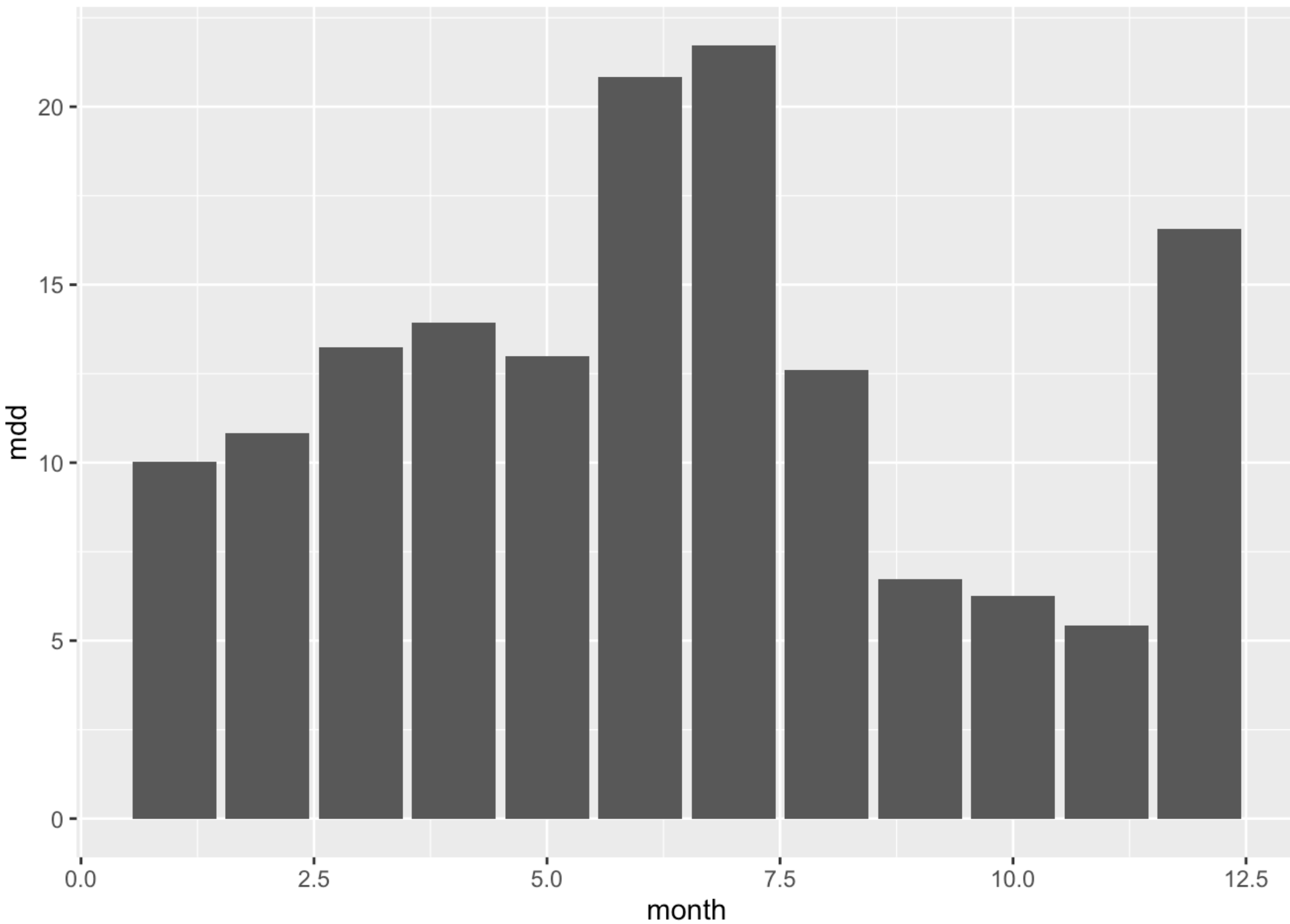
This graph shows that when there is low visibility the number of cancelations on average increases.

Part b)

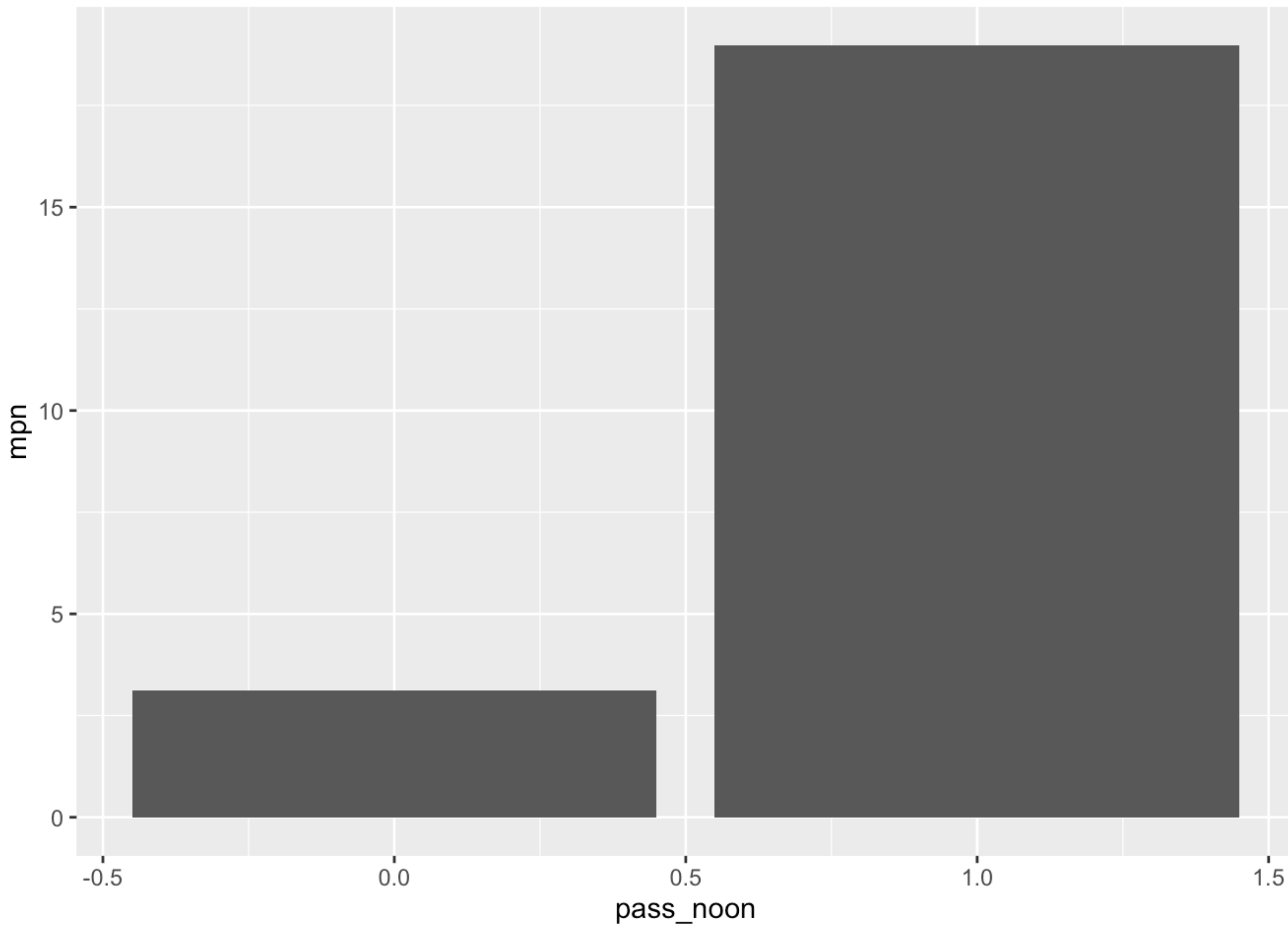
Relationship between Departure Delays and Time

The top 6 months with highes average delay departure are:

```
##    month      mdd
## 7      7 21.72779
## 6      6 20.84633
## 12    12 16.57669
## 4      4 13.93804
## 3      3 13.22708
## 5      5 12.98686
```



Which time of the day has more delay departures? Before noon or after noon?

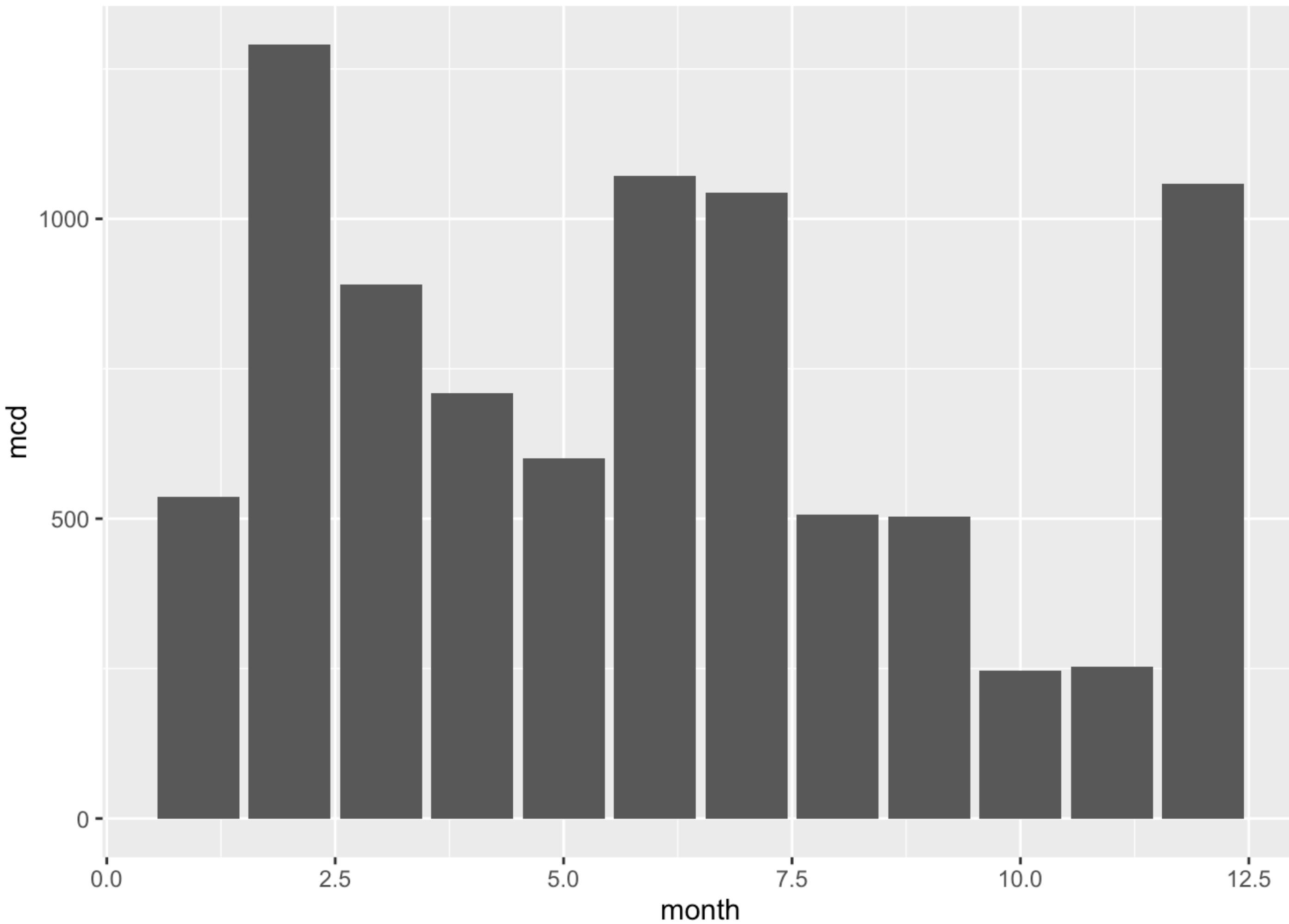


As we can see from the graph above, there are a subtantial larger amount of flights with delay departures after noon. This makes sense, since delaying departure early in the day will mean that its next departure will be more likely to be delayed.

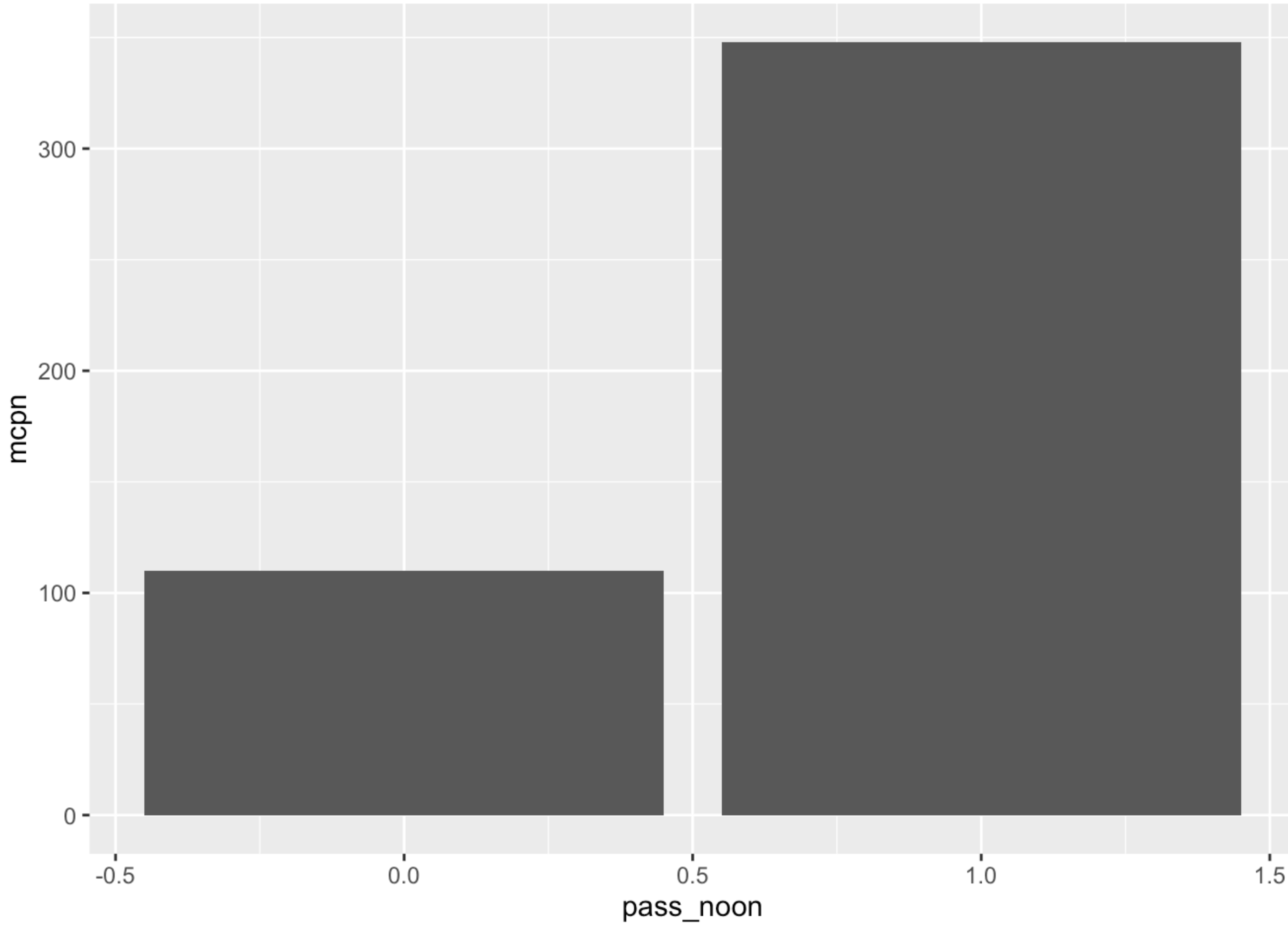
Relationship between Cacelations and Weather

```
##    month  mcd
## 2      2 1291
## 6      6 1072
## 12    12 1059
## 7      7 1043
## 3      3  891
## 4      4  710
```

There is not a big difference between each month and the number of cancelations. But the month with the highest number of cancelations is February.



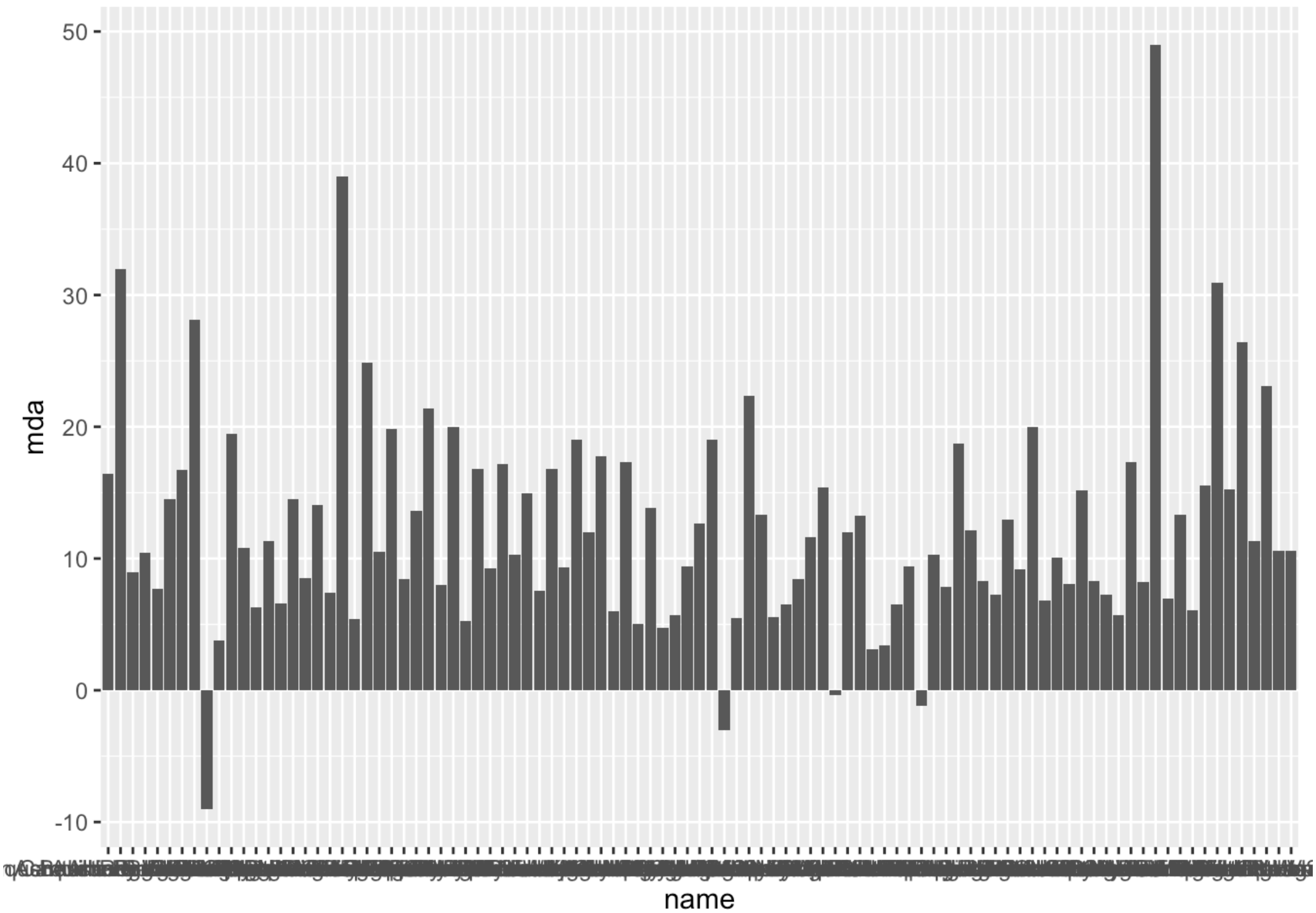
Similarly to departure delays, there are more cancelations past noon.



Part c)

Relationship between Departure Delays and Airport

```
##                      name      mda
## 86       South Bend Rgnl 49.00000
## 20 Columbia Metropolitan 39.00000
## 2            Albany Intl 31.93525
## 91            Tulsa Intl 30.91304
## 8        Birmingham Intl 28.09589
## 93     Will Rogers World 26.42574
```



This graph shows how the average departure delay is distributed within the destination airports. South Bend Reginal is the airport with the highest departure delay.

Relationship between Cacelations and Airports

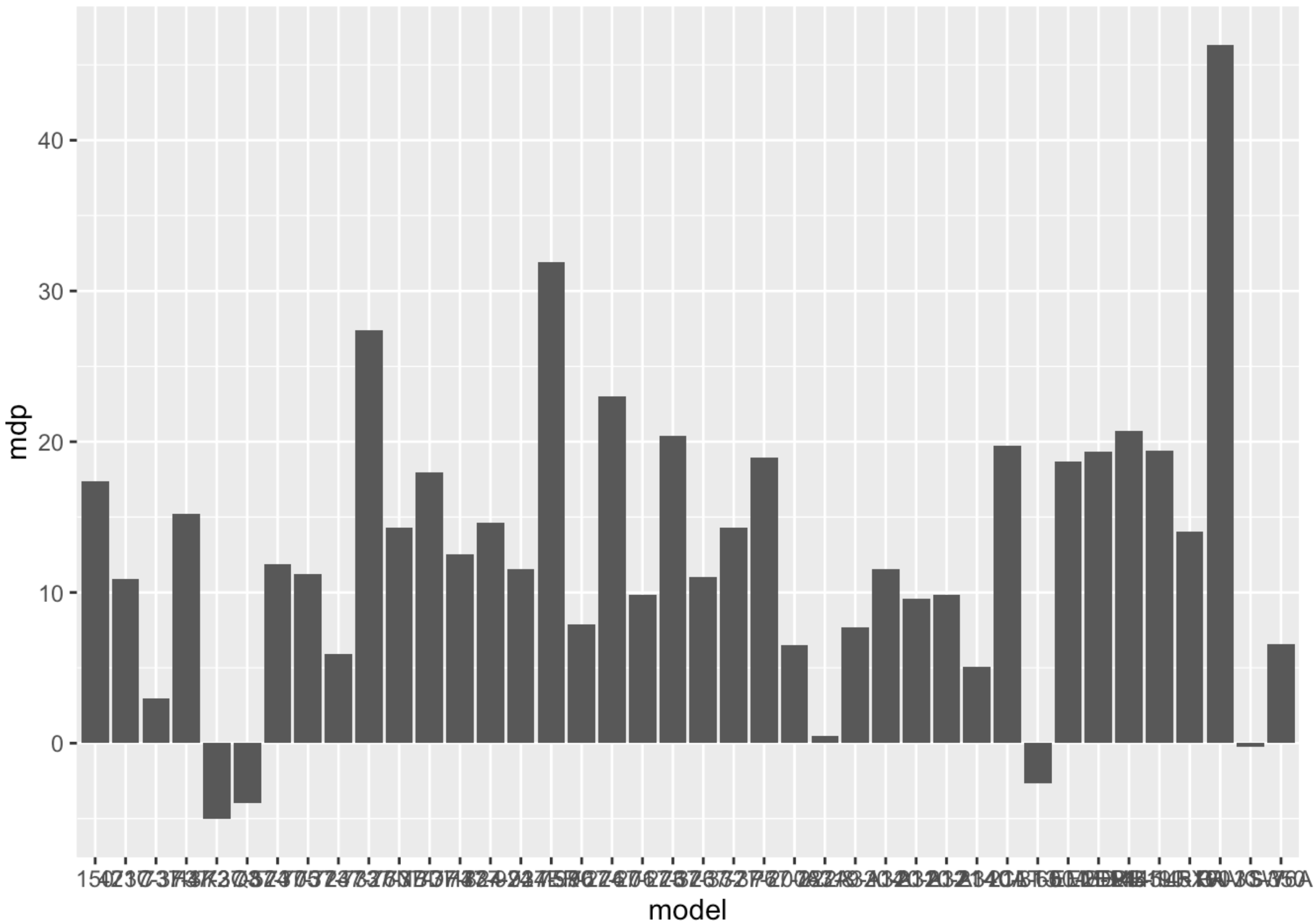
We can see that the airport with the highest number of cancelations is Ronald Reagan Washington.

```
##                                   name mca
## 77      Ronald Reagan Washington Natl 131
## 21             Dallas Fort Worth Intl 123
## 17                 Chicago Ohare Intl 120
## 30 General Edward Lawrence Logan Intl 117
## 75                Raleigh Durham Intl  93
## 15             Charlotte Douglas Intl  88
```

Part d)

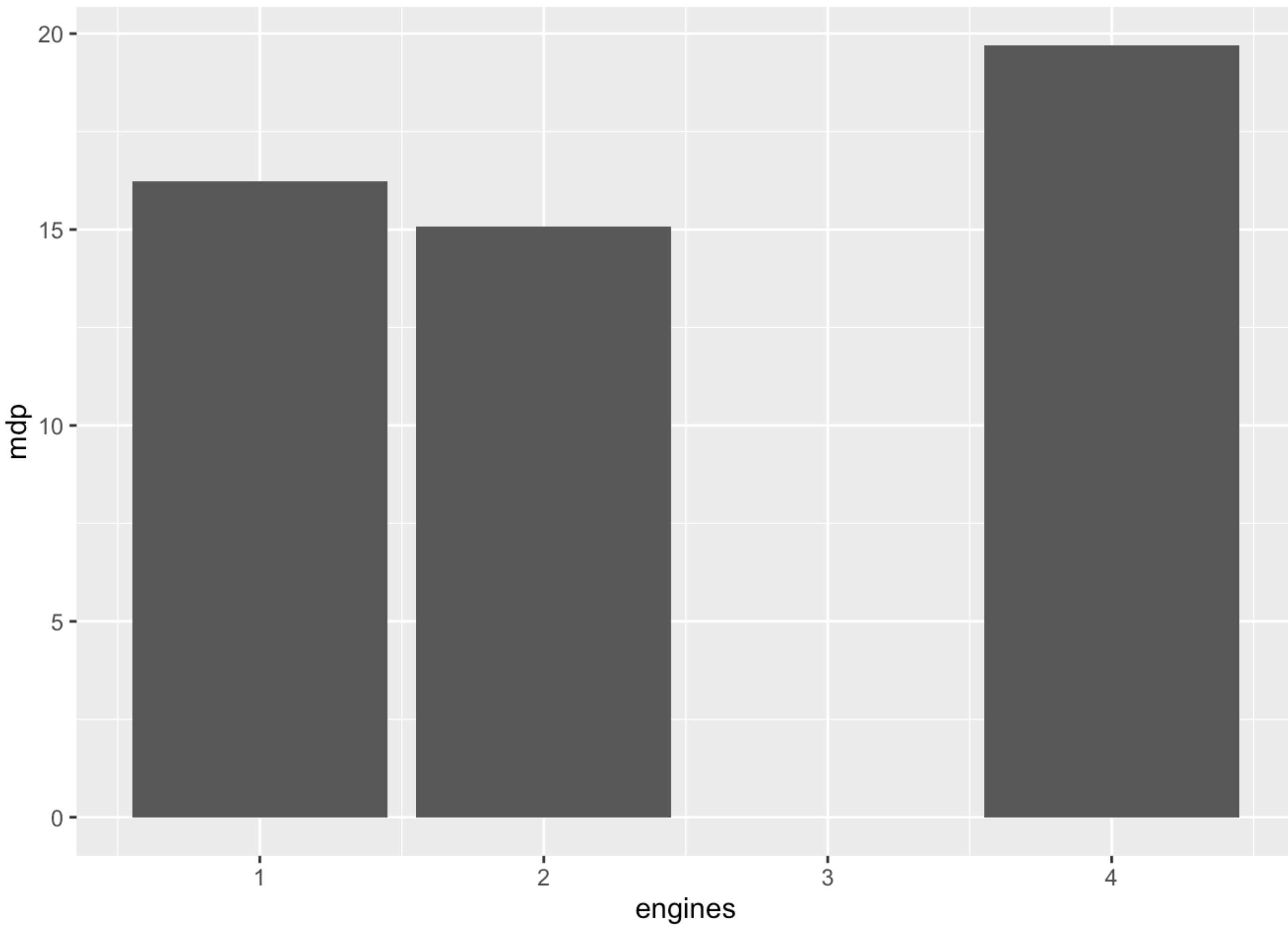
Relationship between Departure Delays and Plane Characteristics

```
##        model      mdp
## 38      G-IV 46.33333
## 16   737-990 31.94444
## 10   737-76N 27.39490
## 18   767-201 23.00000
## 35 EMB-145LR 20.72274
## 20   767-323 20.39130
```



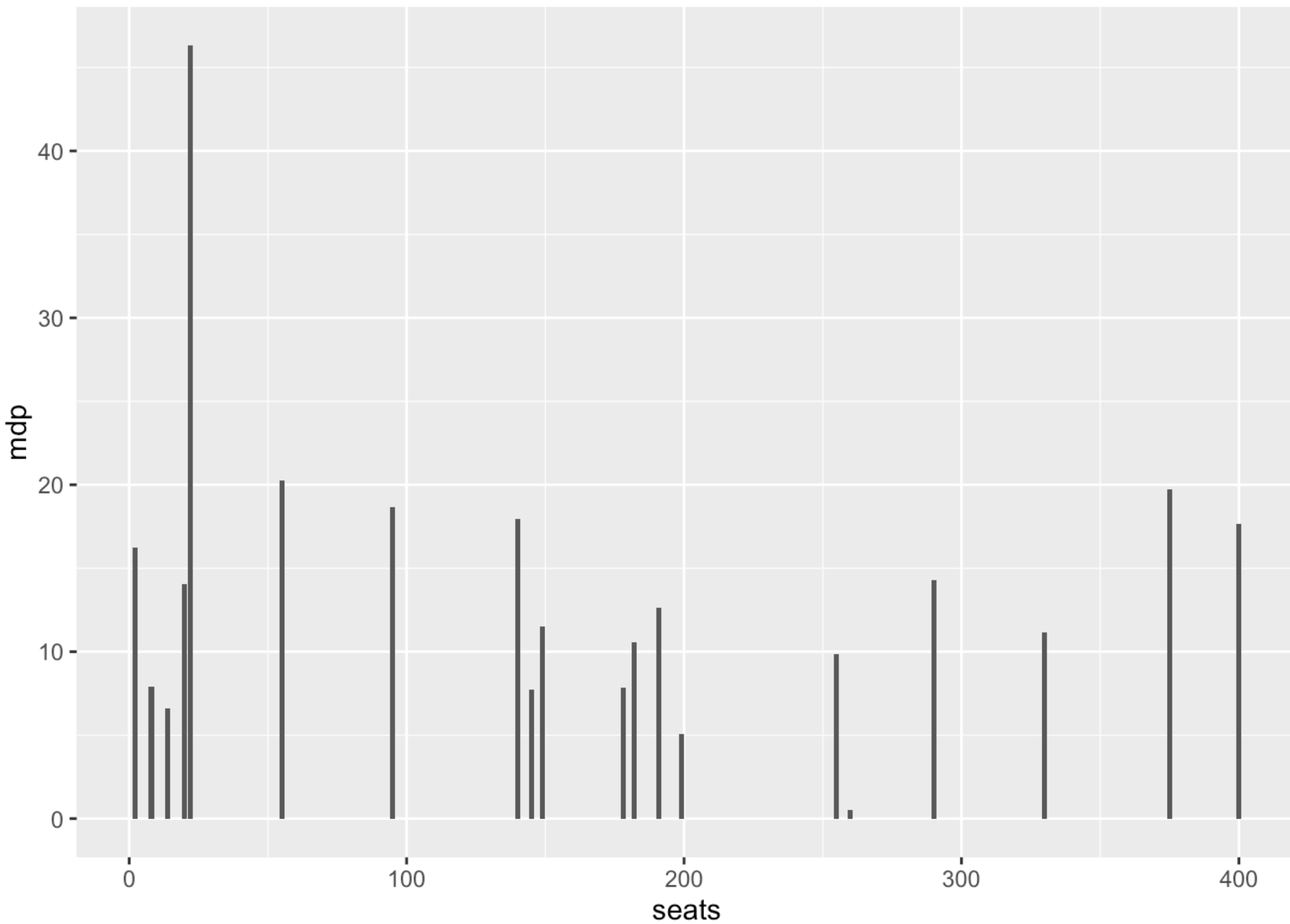
These results show how evenly spread the departure delay is with the model of the plane. It also shows how there are a few models that are extremely high in departure delays, model G-IV, and others that have negative departure delays. Other than those few departure delay seems to be evenly distributed.

```
##   engines      mdp
## 3       4 19.70588
## 1       1 16.24074
## 2       2 15.06339
```



The number of engines in the plane doesn't seem to matter for predicting departure delays.

```
##    seats      mdp
## 5     22 46.33333
## 6     55 20.27353
## 19   375 19.70588
## 7     95 18.66544
## 8    140 17.95130
## 20   400 17.63158
```

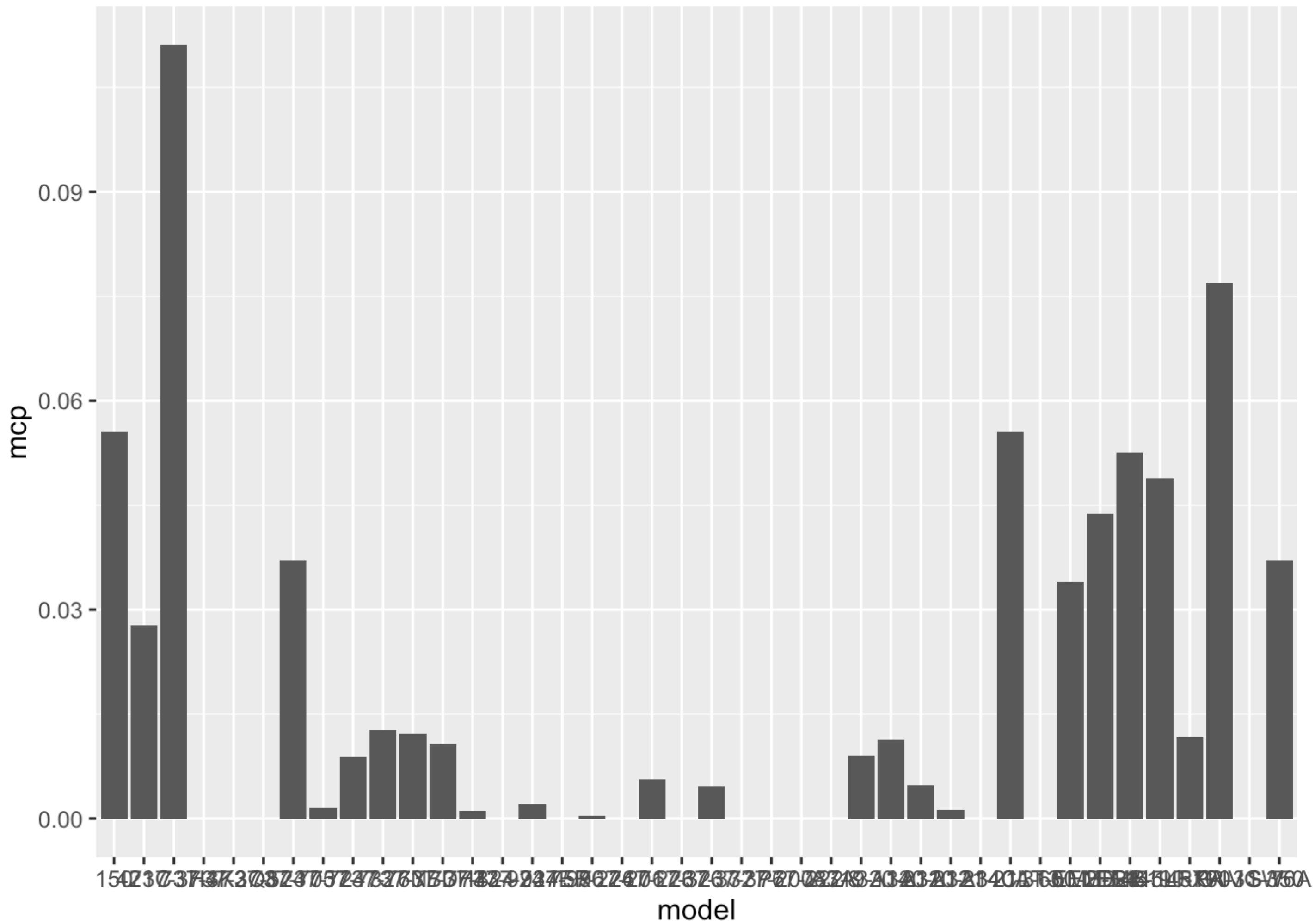


Seats has a similar trait to model. It looks like the airplane model G-IV has 22 seats.

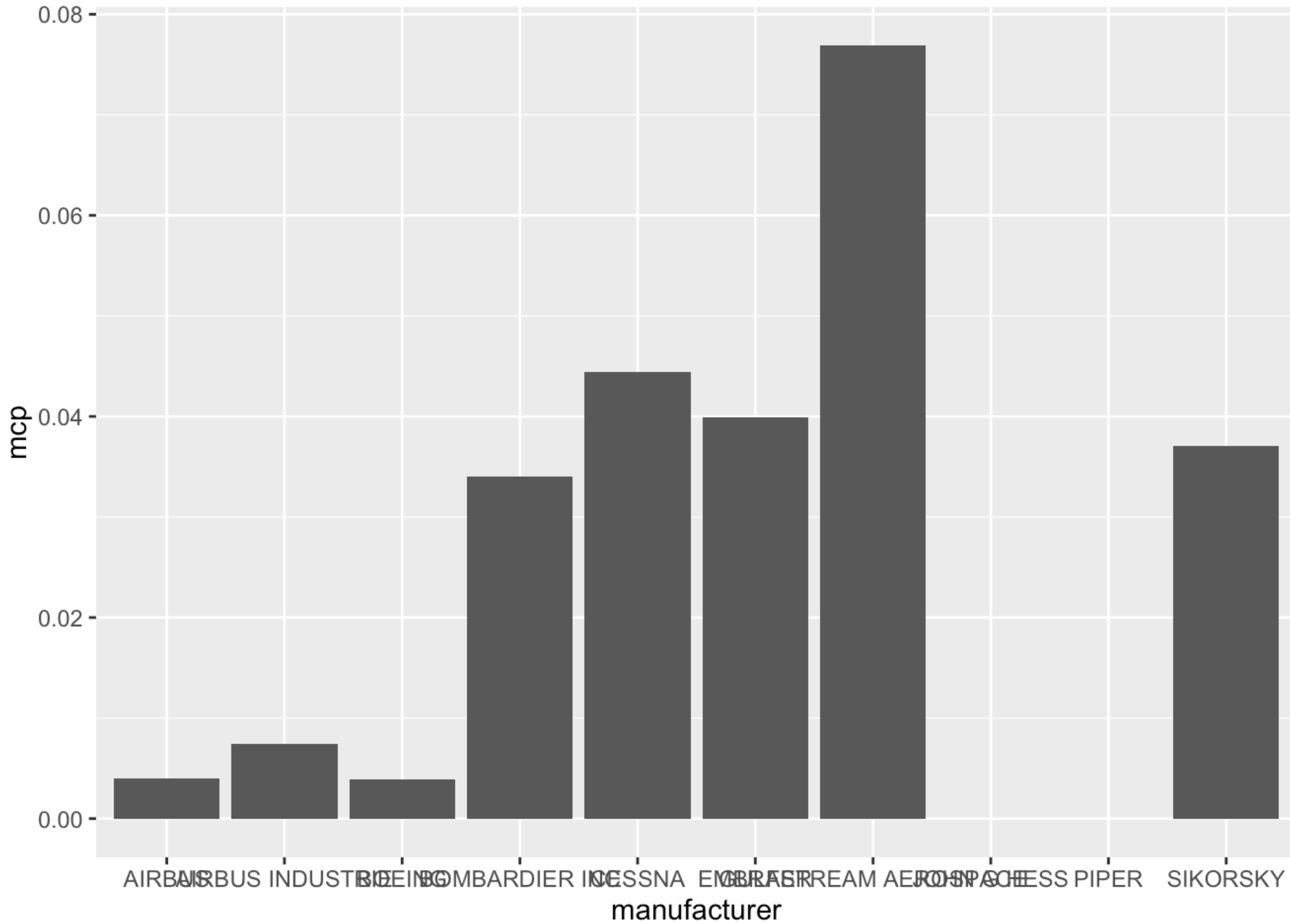
Relationship between Cancelations and Plane Characteristics

```
##        model        mcp
## 3    737-3H4 0.11111111
## 38      G-IV 0.07692308
## 1        150 0.05555556
## 31  A340-313 0.05555556
## 35 EMB-145LR 0.05248839
## 36 EMB-145XR 0.04887677
```

The models with the highest number of flights that were cancelations on average in the EMB-145LR.




```
##            manufacturer         mcp
## 7  GULFSTREAM AEROSPACE 0.076923077
## 5                CESSNA 0.044444444
## 6               EMBRAER 0.039870966
## 10             SIKORSKY 0.037037037
## 4        BOMBARDIER INC 0.033989267
## 2      AIRBUS INDUSTRIE 0.007442322
```


0.08
0.06
0.04
0.02
0.00
mcp
PIPER
SIKORSKY
manufacturer