Analytics Foundations: Problem Set 1

1. The *unc* dataset contains information about UNC graduates from the class of 1984. The data contains a column **salary** and was collected in 1994, at the 10 year class-reunion celebration, in an attempt to measure success of students 10 years post-graduation.

    a. The provost would like you to calculate the average salary 10 years post graduation for the university's new admissions dashboard. *Compute and report that average.*

The average is $241,993 (however, there is a VERY large outlier!!)

    b. Is the provost reporting the appropriate statistic for this dashboard in your opinion? why or why not?

Alas, no! The median is much more representative of this data set.

    c. The provost would also like to include a breakdown of salary by NC residence (column NC_res). He'd like this in the form of a barchart. The following code is a good start, but you might want to change the "fun" parameter (which defines the height of the bars based upon what you noted in part (b.) If you wanted to omit missing values, in the **ggplot()** you would change the dataset from **unc** to **unc[unc$NC_res !='',]**

```
ggplot(unc) +
  geom_bar(aes(x=NC_res,y= salary),
           position = "dodge", stat = "summary", fun = "mean") +
labs(title = "Average Salary by NC Residency") +
  scale_y_continuous(labels = function(x) format(x, scientific =F))
```

Using the median, the salaries appear similar for NC residence versus non-residence.

    d. How would you describe the distribution of the NC_res variable? Does anything about this analysis give you pause before you submit the results to the provost?

There are 4 non-residence for NC and 7 residents for NC and there are 164 missing values!! Would NOT recommend breaking this information down by residency.