STT 811

In-Class Assignment 5

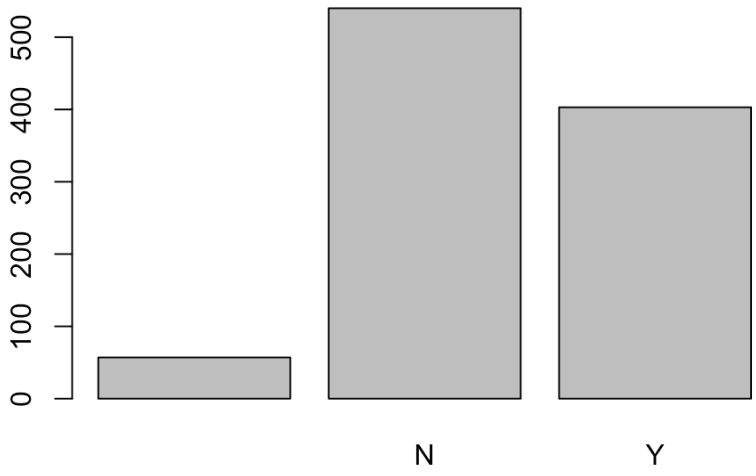
This problem will use the sales_data dataset on D2L. The quantity (this is quantity sold) is the target, and the X1, X2, X3 are input variables.

1. Look first at the X variables. What problems (if any) do you see with these?
    a. The biggest issue we discussed is the missing/empty values from the x2 column, compared to the NAs.
    b. Some explorations:

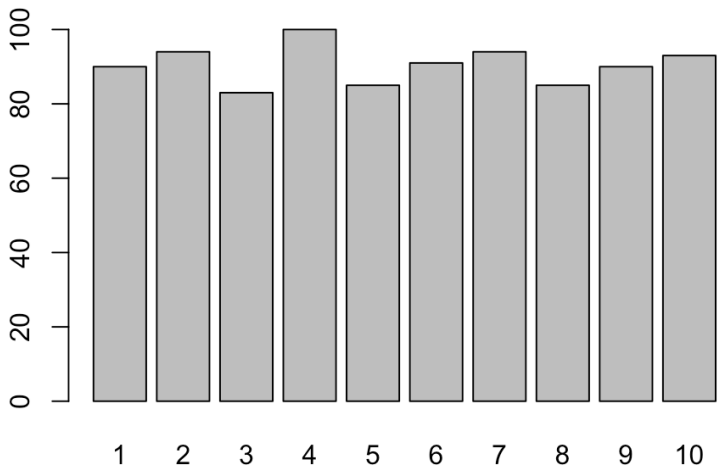
```
> summary(sales$quantity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -93.9   732.6  1008.8  1005.8  1278.9  2206.7
> summary(sales$x1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  79.44   96.63  100.10   99.97  103.51  116.55
> summary(sales$x2)
   Length     Class      Mode
     1000 character character
> summary(sales$x3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   3.000   6.000   5.498   8.000  10.000      95
```

```
> table(sales$x2)

      N   Y
 57 540 403
> table(sales$x3)

  1   2   3   4   5   6   7   8   9  10
 90  94  83 100  85  91  94  85  90  93
```

X2 barplot:



x3: barplot



2. Now look at the target variable. What problem(s) do you see with this one.
    a. The quantity has negative values included, which could be related to returns as mentioned in class.

```
summary(sales$quantity)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-93.9   732.6  1008.8  1005.8  1278.9  2206.7
```

3. Now, clean the dataset by making the target data consistent and clean, and fixing/removing data rows. Describe in a few sentences what you are doing to clean the data.

    I plan to re-categorize all the data with blanks in x2 as its own category to maintain data entries, and NAs in x3 will be set to their own category as well, since x3 seems to fall into buckets, not directly numeric. Also, the units/quantity columns will have to be adjusted to reflect equivalent metrics across the data.