Data Wrangling Steps

Anthony Stachowitz

July 11, 2018

1. Load dataset

This loads the individual datasets.

```
insurance_data <- read_excel("C:/Users/Anthony/Desktop/R studio projects/insurance/insurance_data.xls
x", col_names = FALSE)
```

2. Add headers to the data set.

The data set does not have headers. I will add headers durring this step.

```
names(insurance_data) <- c("Age", "WorkClass", "FnlWgt", "Education", "Education-num", "MaritalStatu
s", "Occupation", "Relationship", "Race", "Sex", "CapitalGain", "CapitalLoss", "HoursPerWeek", "Nativ
eCountry", "Salary")
```

3. Add column to convert less than $50,000 to 1 and more than $50,000 to 0.

```
insurance_data <- mutate(insurance_data, LessThen_50 = ifelse(grepl(">50K", Salary), 0, 1))
```

4. Convert characters to factors

Converted WorkClass, Education, MaritalStatus, and Occupation

```
insurance_data$Occupation <- as.factor(insurance_data$Occupation)
insurance_data$WorkClass <- as.factor(insurance_data$WorkClass)
insurance_data$Education <- as.factor(insurance_data$Education)
insurance_data$MaritalStatus <- as.factor(insurance_data$MaritalStatus)
```

5. Remove the data that is not being used.

Remove columns to make the easier to work with.

```
insurance_data <- insurance_data[, -c(1, 3, 5, 8:14)]
insurance_data
```

```
## # A tibble: 32,561 x 6
##    WorkClass        Education MaritalStatus Occupation Salary LessThen_50
##    <fct>            <fct>     <fct>         <fct>      <chr>        <dbl>
##  1 State-gov        Bachelors Never-married Adm-cleri~ <=50K          1.
##  2 Self-emp-not-inc Bachelors Married-civ-s~ Exec-mana~ <=50K         1.
##  3 Private          HS-grad   Divorced      Handlers-~ <=50K          1.
##  4 Private          11th      Married-civ-s~ Handlers-~ <=50K         1.
##  5 Private          Bachelors Married-civ-s~ Prof-spec~ <=50K         1.
##  6 Private          Masters   Married-civ-s~ Exec-mana~ <=50K         1.
##  7 Private          9th       Married-spous~ Other-ser~ <=50K         1.
##  8 Self-emp-not-inc HS-grad   Married-civ-s~ Exec-mana~ >50K          0.
##  9 Private          Masters   Never-married Prof-spec~ >50K           0.
## 10 Private          Bachelors Married-civ-s~ Exec-mana~ >50K          0.
## # ... with 32,551 more rows
```

6. Check for NA's in the data

```
sapply(insurance_data,function(x) sum(is.na(x)))
```

```
##      WorkClass      Education MaritalStatus     Occupation         Salary
##              0              0              0              0              0
##    LessThen_50
##              0
```