Data Selection Description

Joshua Ferris

8/21/2020

The number of observations

```
nrow(df)
```

```
## [1] 11041
```

The Beach Tan data set has 11,041 observations.

The number of variables

```
ncol(df)
```

```
## [1] 12
```

The Beach Tan data set has 12 variables.

The data type of each variable

Some of the variables were typed incorrectly and so I will convert them to factors.

```
df$UIDStoreLocation = factor(df$UIDStoreLocation)
df$MembershipType = factor(df$MembershipType)
df$MembershipLevel = factor(df$MembershipLevel)
str(df)
```

```
## 'data.frame':    11041 obs. of  12 variables:
##  $ UIDClient       : int  597 17873 26441 31132 31382 44204 50652 81049 98231 9020 ...
##  $ UIDStoreLocation: Factor w/ 10 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ Gender          : Factor w/ 4 levels "","#NULL!","0",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ DateJoined      : Factor w/ 2840 levels "","1/1/2008",..: 1116 377 1171 1171 1599 2150 1918 1185 1
##  $ DaysSinceJoined : int  1694 1855 2785 2785 921 2689 2351 2053 1591 1160 ...
##  $ MembershipType  : Factor w/ 3 levels "0","1","2": 3 3 3 3 3 3 3 3 3 3 ...
##  $ MembershipLevel : Factor w/ 5 levels "0","1","2","3",..: 1 1 2 1 1 1 1 5 1 2 ...
##  $ Age             : int  38 48 76 40 20 20 52 40 63 21 ...
##  $ UVTans          : int  189 0 265 327 90 162 2 34 11 306 ...
##  $ SunlessTans     : int  4 26 58 18 0 17 1 2 0 0 ...
##  $ UpgradeRevenue  : int  56 25 15 0 8 0 0 0 0 0 ...
##  $ RetailRevenue   : num  76 171 190 400 67 ...
```

- Factor: UIDStoreLocation, Gender, DateJoined, MembershipType, and MembershipLevel.
- Int: UIDClient, DaysSinceJoined, Age, UVTans, SunlessTans, and UpgradeRevenue.
- Num: RetailRevenue.

The levels of each factor

```
levels(df$UIDStoreLocation)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```

```
levels(df$Gender)
```

```
## [1] ""       "#NULL!" "0"      "1"
```

```
head(levels(df$DateJoined))
```

```
## [1] ""          "1/1/2008"  "1/1/2011"  "1/10/2002" "1/10/2003" "1/10/2004"
```

```
levels(df$MembershipType)
```

```
## [1] "0" "1" "2"
```

```
levels(df$MembershipLevel)
```

```
## [1] "0" "1" "2" "3" "4"
```

- UIDStoreLocation: 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10.
- Gender: "“,”#NULL¡‘,”0“, and”1".
- DateJoined: "“,”1/1/2008“, 1/1/2011”, “1/10/2002”, “1/10/2003”, “1/10/2004”, and many more dates.
- MembershipType: 0, 1, and 2.
- MembershipLevel: 0, 1, 2, 3, and 4.

The amount of missingness in the data (DataExplorer)

```
library(DataExplorer)
profile_missing(df)
```

```
##             feature num_missing pct_missing
## 1         UIDClient           0           0
## 2  UIDStoreLocation           0           0
## 3            Gender           0           0
## 4        DateJoined           0           0
## 5   DaysSinceJoined           0           0
## 6    MembershipType           0           0
## 7   MembershipLevel           0           0
## 8               Age           0           0
## 9            UVTans           0           0
## 10      SunlessTans           0           0
## 11   UpgradeRevenue           0           0
## 12    RetailRevenue           0           0
```

The above output tells us that the data set has no missing data. However, from looking at the data manually, I have observed empty characters and “#NULL!” in the Gender column. In the date column there are some empty character string also. We will want to clean these values before we perform any analysis.

High-level overview of the analysis I will conduct

I theorize that the more active a client is in their tanning services, the more retail revenue they will generate. To test this theory I will perform some explanatory analysis on the data using linear regression on the response variable RetailRevenue.