UCONN
SCHOOL OF BUSINESS

UNIVERSITY OF CONNECTICUT SCHOOL OF BUSINESS
OPIM 5503 – DATA ANALYSIS USING R

EXTRA ASSIGNMENT

BY

VENKATA JAGANNATH

For the purpose of this project we will follow the SEMMA (Sample, Explore, Modify, Model and Analyze) approach.

Sample

A sample of the data is not required since we will only be doing a logistic regression on our data but we will have to explore the data further and make necessary changes so as to derive an accurate model.

Explore & Modify

In the explore part of this project, I'm trying to answer the following few questions –

- What effect does the patient's age or doctor's age have on consent?
- Does the gender of the doctor influence a patient's to give consent?
- Do the number of reports to carry or number of doctors to visit have an effect on a patient's consent?
- Is there an effect on location (Urban/Rural) on consent?

We will first remove certain variables from our dataset for ease of developing a model. This can be done using the command below.

```
> Consent_IDs <- consent[,c("Consent","P_PatientID","D_docid","D_Specialty","Referrals.to.doc","no.of.patients")]
> consent <- consent[,-c(2,8,10,17,18)]
```

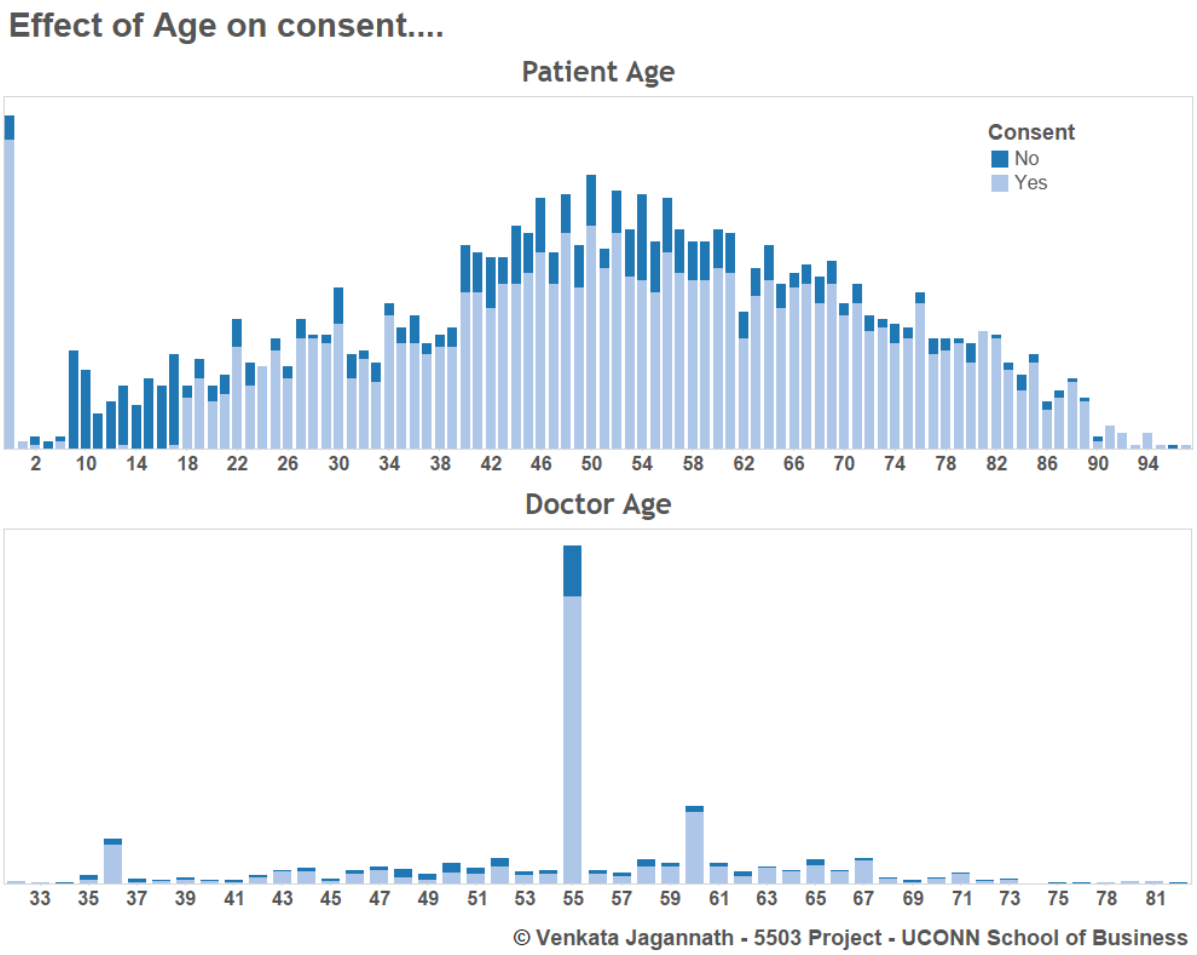
After doing this, we will explore different variables to find patterns and anomalies that need to be fixed before running our model.

We can understand more about the information available on the target variable using –

```
> table(consent$Consent)

   N    Y
 608 2392
```

From the above command we observe that we have more data points that have a consent 'Yes' than those that have consent 'No' i.e we have a class imbalance. We must try to reduce this and bring them to the same level.

Age

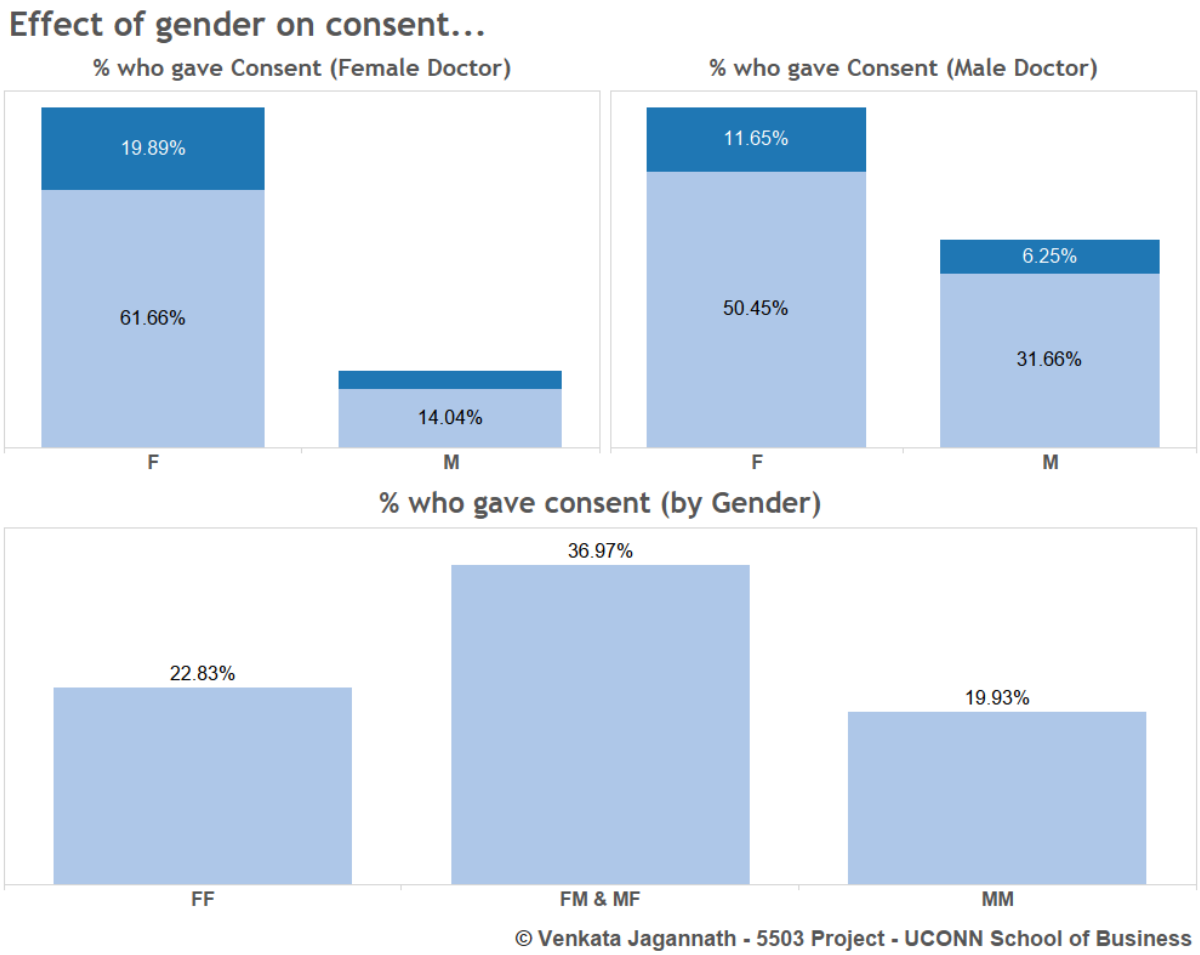


We can clearly notice that there are certain values for Patient Age in our dataset that skew our models. Clearly, we cannot have patients who are zero years old. As we can see from our graph, most records with Age as zero have consented to sharing their information. But, we also know that we have enough information for Consent = 'yes'. Therefore, we remove this data using the commands below.

```
> # Remove values where Age is zero
> sum(consent$P_Age==0) # Number of patients with zero age is 85.
[1] 85
> consent <- consent[consent$P_Age>0,]
```

Gender

We have two different gender variables – Patient gender and Doctor's gender. Lets visualize our data to understand it better -



Insights

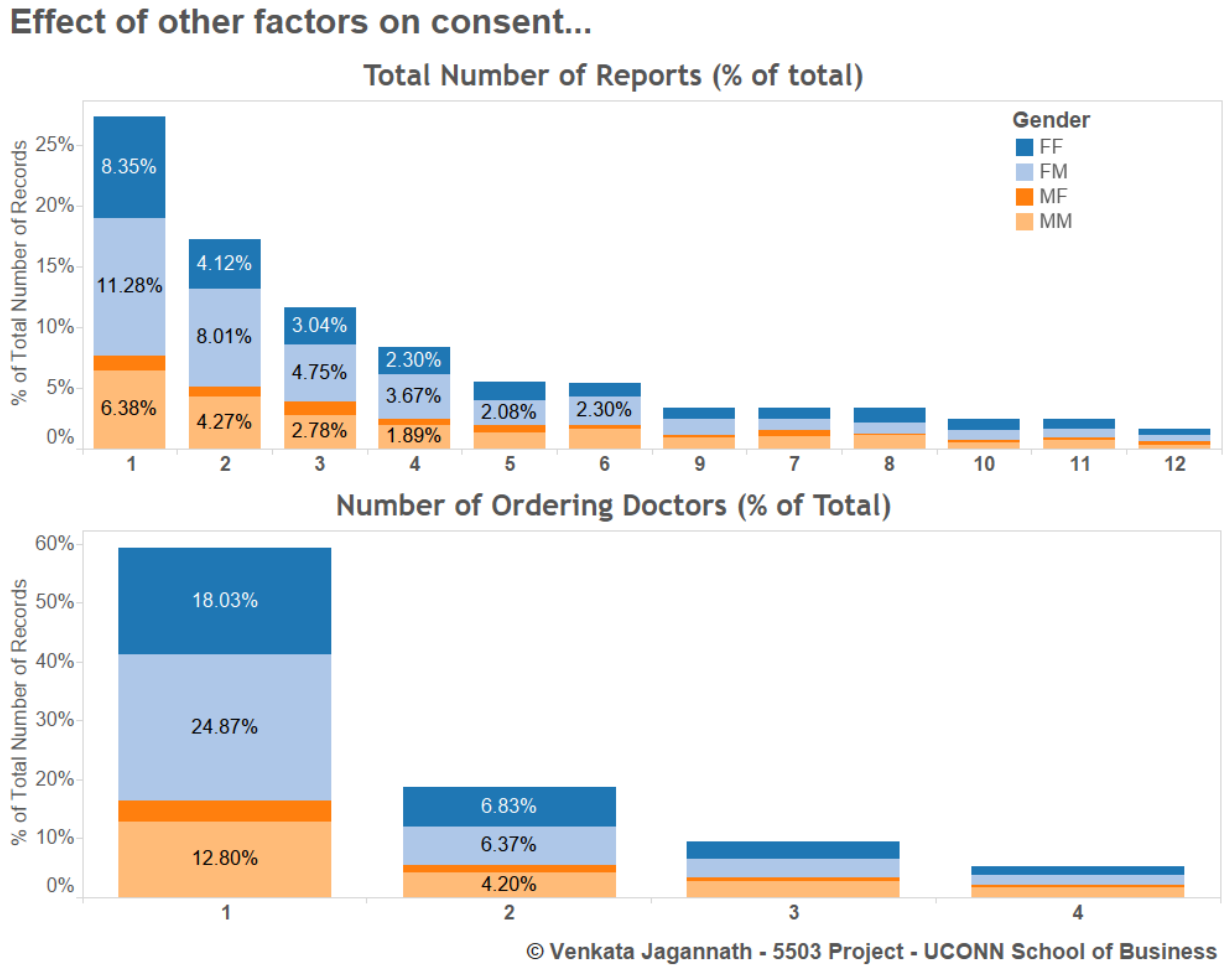
We can observe that more number of women are comfortable with giving consent to a doctor of the same gender. Also, very few men are comfortable with giving consent to a doctor of the opposite sex. The vice versa does not hold true. So, if we do a bit of feature engineering and visualize our new variable 'Gender' we can see that about 36% of our patients are willing to give consent to a doctor of the opposite sex. But about 43% prefer giving consent only to a doctor of the same gender. In these 43%, the men are most unwilling.

For the purpose of reducing dimensionality, we will use the 'Gender' feature going forward and eliminate the other two features using the command below –

```
> # Feature engineering with Patient and Doctor genders
> consent["Gender"]<- paste(consent$P_Gender,consent$D_DocGender,sep="")
> consent <- consent[,-(which(names(consent)%in%c("D_DocGender", "P_Gender")))]
```

No of reports and Number of ordering doctors:

Since we are trying to predict consent for the general population, we can safely assume that most of them will not have more than 12 reports. Even in our data set we can see that about 85 – 90% of patients have less than 12 records. Even the number of ordering doctors is not more than 4 for about 90% of our patients. So, we will remove the outliers and predict for the rest of the information.



Lets look at the R codes to remove these outliers –

```
> sum(consent$P_No.ordering.doctors>4)
[1] 210
> sum(consent$P_No.ordering.doctors>4&consent$Consent=="Y")
[1] 199
```

We can observe that the patients with the number of doctors greater than 4 is 210 and about 199 of those 210 have given consent. But since we have enough data points for consent = “Y”, we will go ahead and remove the data using the command below.

```
> # Reducing the data based on Number of ordering doctors
> consent <- consent[consent$P_No.ordering.doctors<4,]
```

Similarly for the 'Total Number of Reports' field –

```
> sum(consent$P_Total.No..of.Reports>12)
[1] 283
> sum(consent$P_Total.No..of.Reports>12&consent$Consent=="Y")
[1] 244
```

We can observe that 244 out of 283 patients have consent="Y". So, we can remove this information using the command below –

```
> # Reducing the data based on Number of Reports
> consent <- consent[consent$P_Total.No..of.Reports<12,]
```

After making all these changes to our dataset, we can observe the following class breakup for our 'Consent' variable

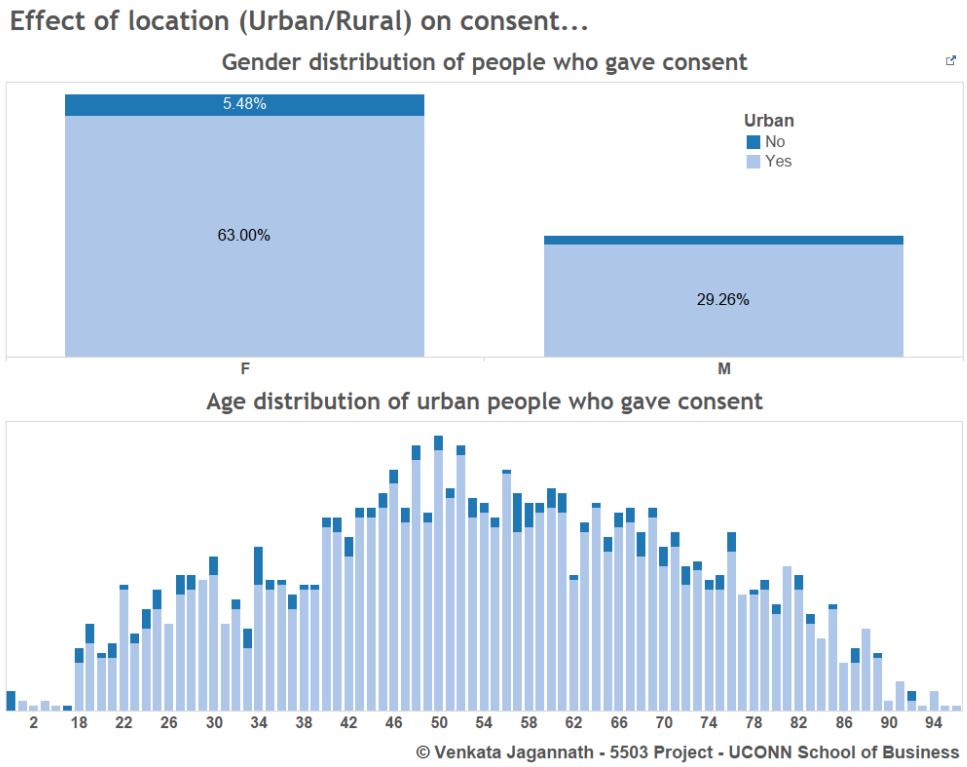
```
> table(consent$Consent)

   N    Y
 541 1693
```

Urban

On this particular variable, we can make several hypothesis –

- Urban people are more unlikely to give consent owing to knowledge of privacy issues.
- However, the contrary is also true wherein people in urban areas are aware of the advantages of sharing their information and are willing to make the trade off between privacy and convenience.
- We can also hypothesize that urban young people are better informed of the recent NSA snooping scandal and are less willing to share their information.

Lets look at all these three hypothesis individually using our chart below.



Hypothesis 1

From the chart in the below part of the graph we can see that this hypothesis is not true since we can see that urban people from all age groups are giving consent.

Hypothesis 2

From both the charts in the visualization above we can say that people irrespective gender are willing to make the privacy & convenience trade off.

Hypothesis 3

We can say that this hypothesis also proves to be incorrect from the graphs above.

Based on the above results, we cannot make any assumption therefore we leave the data unchanged.

Model

We can fit a generalized linear model function which will fit a logistic regression model to our data as shown in the code below –

```
# A generalized regression model is fitted to our data
reg <- glm(consent$Consent~.,data=consent,family = "binomial")
```

The results of this function are as follows –

```
> summary(reg)

Call:
glm(formula = consent$Consent ~ ., family = "binomial", data = consent)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7456   0.1311   0.4918   0.7370   1.6361

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.475e+00  5.088e+00   0.683  0.49462
P_Age                    3.764e-02  3.059e-03  12.306  < 2e-16 ***
P_Total.No..of.Reports   1.405e-02  2.265e-02   0.620  0.53517
P_No.ordering.doctors    9.467e-01  1.203e-01   7.871 3.53e-15 ***
P_Distance               1.682e-03  6.080e-04   2.766  0.00567 **
D_GroupNMD              -1.341e-01  1.884e-01  -0.712  0.47647
D_DocAge                 3.795e-02  7.721e-03   4.916 8.84e-07 ***
zip                     -5.611e-04  3.599e-04  -1.559  0.11901
zpop                    -5.265e-06  3.551e-06  -1.483  0.13812
urban                    6.059e-01  1.844e-01   3.287  0.00101 **
competition              3.254e-04  1.038e-03   0.313  0.75391
GenderFM                 1.833e-01  1.361e-01   1.346  0.17828
GenderMF                -1.410e-01  2.383e-01  -0.592  0.55405
GenderMM                 1.264e-01  1.579e-01   0.800  0.42344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2556.3  on 2347  degrees of freedom
Residual deviance: 2149.4  on 2334  degrees of freedom
AIC: 2177.4

Number of Fisher Scoring iterations: 5
```

We can create a new column for our prediction using the code below and also check our results using the code below –

```
# The prediction is saved to a new column
consent["pred"] <- predict(reg,consent,type="response")

# We use a cutoff of 0.7 to derive our prediction
consent$pred <- ifelse(consent$pred>0.7,"Y","N")

# The function gives all statistics of our prediction
confusionMatrix(consent$Consent,consent$pred)
```

The results of the confusion matrix is as follows –

```
> confusionMatrix(consent$Consent,consent$pred)
Confusion Matrix and Statistics

          Reference
Prediction    N    Y
         N  311  239
         Y  351 1447

               Accuracy : 0.7487
                 95% CI : (0.7307, 0.7662)
    No Information Rate : 0.7181
    P-Value [Acc > NIR] : 0.0004596

                  Kappa : 0.3458
 Mcnemar's Test P-Value : 4.882e-06

            Sensitivity : 0.4698
            Specificity : 0.8582
         Pos Pred Value : 0.5655
         Neg Pred Value : 0.8048
             Prevalence : 0.2819
         Detection Rate : 0.1325
   Detection Prevalence : 0.2342
      Balanced Accuracy : 0.6640

       'Positive' Class : N
```

Analyze

We can see that we are able to predict consent correctly only about 75% of the time. Our sensitivity and specificity are 46% and 85% respectively. We have a kappa value of 0.34.

This model has a few Type-1 and Type-2 errors and it can be improved with a little more information on the patient.

Future course of action

- A variable with the education level of the patient can be useful in making a better prediction
- Better use of some of the other variables such as Distance, PatientID and DocId can help give us a better model.
- A categorization of gender specific specialty such as pregnancy for women can provide a better model.