DOM 207: IDSBRP
MINI PROJECT 1
HOW TO IMPROVE A RESTARAUNT 101

COLLABORATORS:

1. Sanjana Nair - 2010110764
2. Aditya Arora - 2010110038

INSTRUCTOR: Dr. Jaideep Ghosh

INTRODUCTION:

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts," Sherlock Holmes proclaims in Sir Arthur Conan Doyle's *A Scandal in Bohemia.*

Data processing and Analyzing is a field that has grown exponentially in recent years. Almost everywhere you go, you soon understand the importance of data and how to use the data for various inferences. This course has been helping us learn how to efficiently use the data available to us and we are excited to have an opportunity to try and test our knowledge.

We've used the given "Serve.xlsx" excel sheet given to us and only after thoroughly cleaning the data are using it to make various inferences.

Here we have presented the data that we have studied by visualization and analysis. Using this we have come up with plenty of ideas to make sure our restaurant flourishes.

DATA CLEANING:

We started by downloading the excel sheet uploaded and saving it also as a CSV file and then uploaded the said CSV file in our python program (Named the imported CSV file as df).

DEALING WITH NULL VALUES-

Using the print(df.isnull().sum()) code we get a list of the number of null values in each column. According to the respective number of null values, each column contains we have filled the data with appropriate entries.

```
Amount      0
Tip         0
Gender      0
Smoker      7
Day         2
Time        0
Partysize   1
```

Using print(df.mode()) code we get to know what is the most recurring value in each column.
Using print(len(df)) code we get to know the database has 327 rows i.e. 327 entries in each column.

SMOKER-

With the smoker column, we see that the number of null values is 7 and the more recurring entry in the column is NO. Since out of 327 entries only 7 are null values and the number of No are more we decided to fill the null values with No as that doesn't affect the database.

So using df["Smoker"].fillna("No",inplace=True) code we fill null values in "Smoker" with NO.

DAY-

Similarly, with Day we get to know the most recurring value is Saturday and since the number of null values is just 2, We fill the null values with Saturday using
df["Day"].fillna("Saturday",inplace=True) code.

PARTYSIZE-

For partysize, we see there's only one null value and since the mode and median for partysize is 2, we fill the null value with 2 using the code -

```
median = df['Partysize'].median()
df['Partysize'].fillna(median, inplace=True)
```

Thus we replace all the null values with the appropriate entries in all columns.

DEALING WITH INCOMPLETE VALUES-

We noticed that some of the entries though not null values are incomplete and decided to rectify this.

AMOUNT

In the Amount column there are values where instead of the decimal point there is a “-” or “,” present, to fix this we use the following code.

```
df['Amount'] = df['Amount'].str.replace('-','.')
df['Amount'] = df['Amount'].str.replace(',','.')
```

TIME

In the Time column, there are values with both LD entered and to, later on, split this into two rows with lunch and dinner respectively we first insert a comma which will be our separator between the two letters using

```
df['Time'] = df['Time'].str.replace('LD','L,D')
```

Since they are now separated with a comma we can use the following code to split it into two rows

```
df = df.assign(Time=df.Time.astype(str).str.split(',')).explode('Time').reset_index(drop=True)
```

ONCE we are done with this we decided to CONVERT our file back to excel to edit the excel workbook with all necessary entries

```
df.to_excel("ServeNEWPROJECT.xlsx", sheet_name="1", index=False)
```

We completed all the incomplete entries. Since we had S and San entries and the mode of Day is Saturday we replaced these entries with Saturday.

So now we can use the following code to replace any incomplete entry:

```
replacement_pair={"M":"Male","Mal":"Male","F":"Female","N":"No","Y":"Yes","D":"Dinner","L":
"Lunch","Fri":"Friday","Sat":"Saturday","Sun":"Sunday","Thur":"Thursday","Thurs":"Thursday","S"
:"Saturday","San":"Saturday","LD":"L,D"}
wb=openpyxl.load_workbook("ServeNEWPROJECT.xlsx")
for ws in wb.worksheets:
    for row in ws.iter_rows():
        for cell in row:
            if cell.value in replacement_pair.keys():
                cell.value=replacement_pair.get(cell.value)
wb.save("DOM207NEWSERVECLEAN.xlsx")
```

NOW we have our fully clean database ready under the excel file name DOM207NEWSERVECLEAN.xlsx to go ahead with our statistical analysis and visualization.

DEALING WITH OUTLIER DATA-

Once in excel, in our sheet named 1(with outlier) we use the filter function to check all the entries in all columns. In excel itself we deal with outlier data.

Since the amount and tip are personal to the customer there is no outlier data in these two columns. However, in the partysize column using sort (highest to lowest) we notice there are outlier data.

SO,

USING QUARTILE FUNCTION ON PARTYSIZE TO FIND OUTLIERS.

QTL 1	2
QTL 3	3
IQR	1
Lower limit	0.5
Upper Limit	4.5

Once we do this we create a new column called outlier and use the OR function to enter True for all outlier data and False for non - outlier data. Once this is done we filter out the outlier data and copy-paste the remaining data into a new sheet naming it as (2(without outlier)).

BUSINESS PROPOSAL:

With the data visualization and analysis stated later on we have come up with the following proposals to help increase our restaurant's profits and popularity.

To split the restaurant into two floors with the above floor being a Smoke friendly area (with balconies and proper ventilation) at the mere extra charge of 3/- (14% of the mean of Amt smoker)

(USING 1.1 AND 2.1)

The t-test between the columns amount smoker and amount nonsmoker indicates that mean of smoker parties is greater than of nonsmoker parties, and the t-test between smoker party size and nonsmoker party size tells us the mean of smoker party size is lesser than of nonsmoker party size which suggests a higher willingness to pay by smoker customers even after having a smaller party size. Moreover, the graph tells us that even though nonsmoker parties are more there is a considerable number of smoker parties, So to increase the number of smoker parties having a more comfortable setting for them to smoke, drink and eat on a separate floor will benefit the restaurant.

Sunday buffets during lunch along with a kids-friendly ala carte'

(USING 1.2,1.3,1.5,1.6 AND 2.2,2.3)

The t-test between the columns' party size sunday and party size on other days provides that the mean of the sunday party sizes is slightly higher than the party sizes on other days. However, graph 1.6 suggest that the number of people coming on sunday, represented by the number of parties that come on sunday, is quite low. So increasing the number of people on Sunday by introducing a buffet can lead to a much higher increase in amounts and partysize on sunday as compared to saturday leading to higher profits for the restaurant. The reason why we are not introducing a buffet for Sunday dinner is that the tips (1.5) and the amount we make on sunday dinner are already high and introducing a buffet will disrupt that. Moreover, a kids-friendly ala carte for children who will not utilize the buffet will benefit the parents and us as it gives a further incentive for more people to come into the restaurant. Since (Using 2.3) we see the mean of Partysize 3 AND 4 is higher than Partysize 1 AND 2 AND the number of Partysize 1&2 entries is more than Partysize 3&4, having a buffet gives the incentive to get more profits.

Couple's lunch date special meals

(USING 1.3,1.4 AND 2.4)

Using 1.3 We see there is a major peak in the number of partysize 2 people. Moreover, with 1.4 we see that Partysize 2 people occasionally spend a considerable amount of money. So, Assuming that majority of this would be couples incoming if we introduce a COUPLE SPECIAL menu of slightly higher prices and special items we can increase the profit we make. The t-test between the columns suggests that the mean of the AMT THUR is lower even despite having such a high Partysize 2 intake (1.3).

Weekend starter (Fridays) bash- 15% off on Drinks and Desserts

(USING 1.2,1.3)

Seeing with 1.2 and 1.3 we see how Friday for some reason has less number of people coming in and less amount of money we are making. So by introducing the above idea we can considerably increase the number of people who come to our restaurant without affecting our profits much.

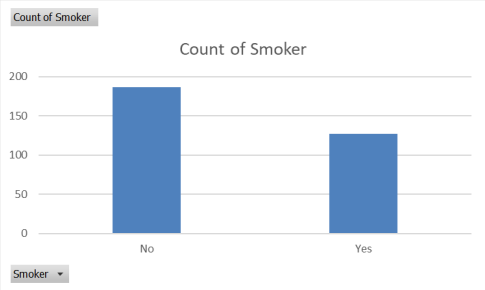
1] DATA VISUALIZATION:

We use excel itself to visualize our data in various combinations.

The excel sheet we have presented contains a lot of sheets as we segregated values for various t-testings using R later on. (For reference - AMT-AMOUNT and PS-PARTYSIZE)

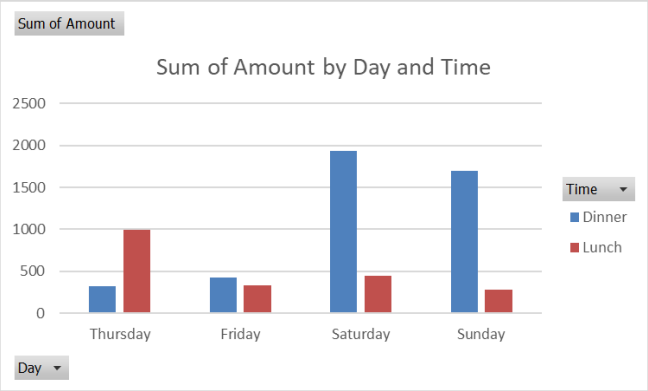
1.1 USING SMOKER SHEET

In this we just infer the number of smoker and nonsmoker parties we have. We notice that even though the number of nonsmoker parties is higher there is still a considerable amount of intake of smoker parties and will be using the information for a business strategy later on.



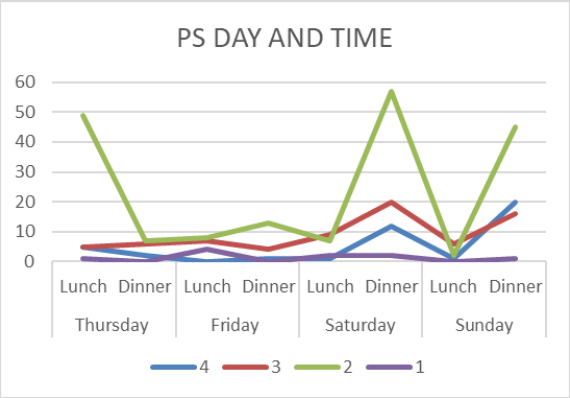
1.2 USING AMT DAY TIME SHEET

In this sheet we visualize the range of how much amount we receive depending on the Day and Time. Once we get the graph we notice how our intake of money is high on Saturday and Sunday during dinner time. We also notice a slight high on Thursdays during lunch.



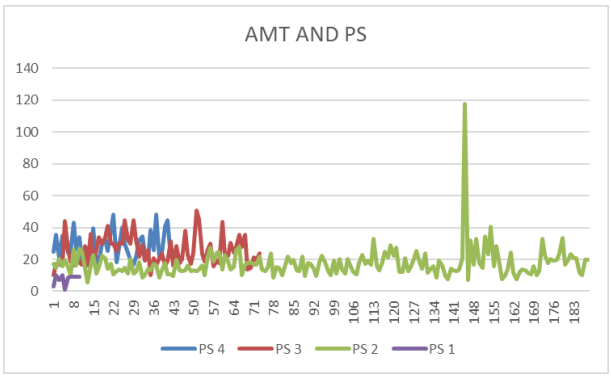
1.3 USING PS DAY TIME SHEET

In this sheet we learn how the intake of particular partysizes varies with each time and day. So here we see how PS 2 has a spike during Thursday Lunch, Saturday Dinner, and Sunday Dinner. Also how there is a general high intake during the weekend mainly for dinner.



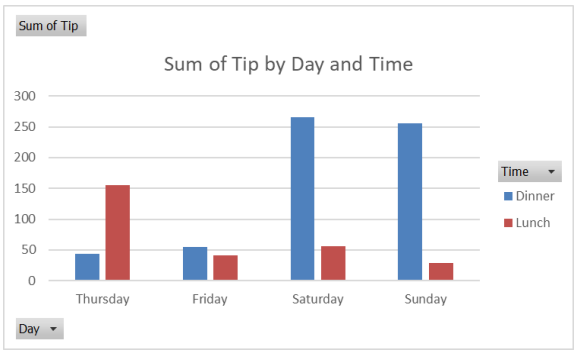
1.4 USING AMT PS SHEET

In this we plot a line graph just to try and understand how much each party pays for their meal. We infer what is naturally expected that apart from a handful of exceptional cases we notice that with the increase in partysize (3,4) there is a higher intake of amount from them.



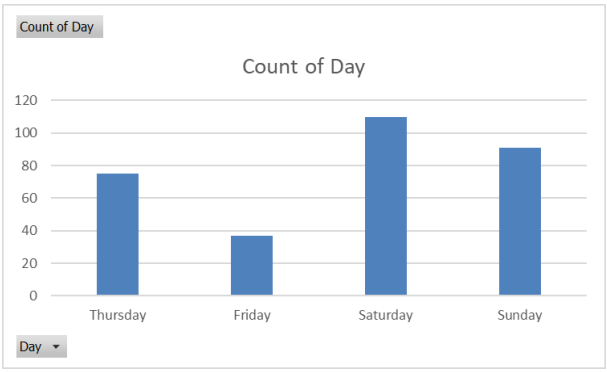
1.5 USING TIP DAY TIME SHEET

In this graph we notice that there's an increase in the number of tips that our restaurant's waiters receive during the Weekend. This is majorly due to the increase in the number of people who come to our restaurant during the weekend.



1.6 USING DAY SHEET

To see the intake of entries depending on the day of the week.



2] DATA ANALYSIS IN R:

Over R during our t-testing, our null hypothesis is that the means for the given variables are equal and are testing for the same and if the p-value is less than 5% (alpha) then we can safely reject the null hypothesis.

2.1 USING ONE SIDED TWO SAMPLE T-TESTS - AMT-SMOKER&NONSMOKER AND PS SMOKER&NOSMOKER

Here we see the mean of AMT SMOKER **>** AMT NONSMOKER and PS SMOKER **<**PS NONSMOKER and we see that they are not equal.

```
> library("readxl")
> #2.1
> DFAS<-read_excel("C:\\Users\\sanjana\\OneDrive\\Desktop\\DOM207MINIPROJECT1-SANJANA ADITYA\\DOM207NEWSERV
ECLEAN.xlsx",sheet = "AMT SMOKER")
> a=DFAS[["AMT SMOKER"]]
> b=DFAS[["AMT NON SMOKER"]]
> t.test(a, b, alternative = c("greater"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  a and b
t = 1.8638, df = 311, p-value = 0.03165
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2572078       Inf
sample estimates:
mean of x mean of y
 21.87890  19.63914


> DFPS<-read_excel("C:\\Users\\sanjana\\OneDrive\\Desktop\\DOM207MINIPROJECT1-SANJANA ADITYA\\DOM207NEWSERV
ECLEAN.xlsx",sheet = "PS SMOKER")
> c=DFPS[["PS SMOKER"]]
> d=DFPS[["PS NON SMOKER"]]
> t.test(c, d, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  c and d
t = -1.7612, df = 311, p-value = 0.03959
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
         -Inf -0.009764203
sample estimates:
mean of x mean of y
 2.377953  2.532258
```

2.2 USING ONE-SIDED TWO SAMPLE T-TESTS- PS-DAY

Here we see the mean of PS THURSDAY < PS SUNDAY, PS FRIDAY < PS SUNDAY and PS SATURDAY < PS SUNDAY.

```
> #2.2
> DFPD<-read_excel("C:\\Users\\sanjana\\OneDrive\\Desktop\\DOM207MINIPROJECT1-SANJANA ADITYA\\DOM207NEWSERV
ECLEAN.xlsx",sheet = "PS DAY")
> h=DFPD[["PS THURS"]]
> i=DFPD[["PS FRI"]]
> j=DFPD[["PS SAT"]]
> k=DFPD[["PS SUN"]]
> t.test(h,k, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  h and k
t = -3.125, df = 164, p-value = 0.001052
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1752313
sample estimates:
mean of x mean of y
 2.320000  2.692308
```

```
> t.test(i,k, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  i and k
t = -2.8869, df = 126, p-value = 0.00229
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1913106
sample estimates:
mean of x mean of y
 2.243243  2.692308
```

```
> t.test(j, k, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  j and k
t = -2.0382, df = 199, p-value = 0.02142
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.04327138
sample estimates:
mean of x mean of y
 2.463636  2.692308
```

2.3 USING ONE-SIDED TWO SAMPLE T-TESTS- AMT-PS

Here we see the mean of AMT PS1< AMT PS2 and AMT PS3 < AMT PS4. Here NULL hypothesis can be rejected.

```
> #2.3
> DFAP<-read_excel("C:\\Users\\sanjana\\OneDrive\\Desktop\\DOM207MINIPROJECT1-SANJANA ADITYA\\DOM207NEWSERV
ECLEAN.xlsx",sheet = "AMT PS")
> p=DFAP[["PS 1"]]
> q=DFAP[["PS 2"]]
> r=DFAP[["PS 3"]]
> s=DFAP[["PS 4"]]
> t.test(p,q, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  p and q
t = -3.2821, df = 196, p-value = 0.00061
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -4.910672
sample estimates:
mean of x mean of y
  7.53900  17.43048
```

```
> t.test(r,s, alternative = c("less"), mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  r and s
t = -1.9681, df = 113, p-value = 0.02575
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.5375272
sample estimates:
mean of x mean of y
 25.52767  28.94429
```

2.4 USING ONE-SIDED TWO SAMPLE T-TESTS AMT-DAY

Here we see how the mean of AMT THUR < AMT FRI, AMT THUR < AMT SAT, AMT THUR < AMT SUN.

```
> #2.4
> DFAD1<-read_excel("C:\\Users\\sanjana\\OneDrive\\Desktop\\DOM207MINIPROJECT1-SANJANA ADITYA\\DOM207NEWSER
VECLEAN.xlsx",sheet = "AMT DAY")
> x=DFAD1[["AMT THUR"]]
> y=DFAD1[["AMT FRI"]]
> z=DFAD1[["AMT SAT"]]
> w=DFAD1[["AMT SUN"]]
> t.test(x,y, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  x and y
t = -1.8269, df = 110, p-value = 0.03521
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.2661069
sample estimates:
mean of x mean of y
 17.60520  20.49784

> t.test(x,z, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  x and z
t = -2.3767, df = 183, p-value = 0.009249
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.213542
sample estimates:
mean of x mean of y
 17.60520  21.59164

> t.test(x,w, alternative = c("less"),
+        mu = 0, paired = F, var.equal = T , conf.level = 0.95)

        Two Sample t-test

data:  x and w
t = -3.2552, df = 164, p-value = 0.0006883
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -2.029666
sample estimates:
mean of x mean of y
 17.60520  21.73198
```

CONCLUSION:

With the data available to us we believe that we have thoroughly analyzed our data and given the best possible solutions for our restaurant's success. We hope you find our work efficient and agree to our proposals for our company to flourish.