Investigate Possible Reasons for Appointment No-Shows

Data Sample: Medical Facility in Brazil

Table of Contents

Introduction

Using data from over 100,000 medical appointments in Brazil, analysis is being done below to hopefully uncover reasons as to why some patients do not show up at their scheduled day/time. When this happens, it means others potentially miss out on receiving medical care or those needing continuous treatment aren't receiving it. The following questions may determine some causes for no-shows to help prevent them in the future.

1.) Do SMS confirmations minimize the likelihood of a no-show?
2.) How many of the overall no-shows are appointments for patients who have a non-handicap medical condition?
3.) Does substance use in the form of alcoholism account for a bigger number of no-shows?

Import/Setup

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%pylab inline
```

```
Populating the interactive namespace from numpy and matplotlib
```

Data Wrangling

General Properties

In [2]:
```
#Downloaded noshowappointments-kagglev2-may-2016.csv from https://s3.amazonaws.com/vide

#Read the CSV file in Jupyter by turning into a data frame
```

```
brazilmed_df=pd.read_csv(r'C:\Users\David\Dropbox\Udacity Data Analyst Nanodegree\Proje

#Do some preliminary observation of the data structure
brazilmed_df.head()
```

Out[2]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scho
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	

In [3]:
```
#Inspect data frame for null values
brazilmed_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   PatientId       110527 non-null  float64
 1   AppointmentID   110527 non-null  int64  
 2   Gender          110527 non-null  object 
 3   ScheduledDay    110527 non-null  object 
 4   AppointmentDay  110527 non-null  object 
 5   Age             110527 non-null  int64  
 6   Neighbourhood   110527 non-null  object 
 7   Scholarship     110527 non-null  int64  
 8   Hipertension    110527 non-null  int64  
 9   Diabetes        110527 non-null  int64  
 10  Alcoholism      110527 non-null  int64  
 11  Handcap         110527 non-null  int64  
 12  SMS_received    110527 non-null  int64  
 13  No-show         110527 non-null  object 
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
```

No null values present.

Data Cleaning

In order to understand what factors might be impacting patients who no-show their appointments, I need to make some of the data easier to work with.

In [4]:
```
#Change AppointmentDay column from strings to datetime objects
brazilmed_df['AppointmentDay'] = pd.to_datetime(brazilmed_df['AppointmentDay'])
```

In [5]:
```
#For the purposes of what I plan to investigate, the following variables can be omitted
#PatientIdGender,ScheduledDay,Age,Scholarship,Handcap
```

```
brazilmed_df.drop(['PatientId','Gender','ScheduledDay','Age','Scholarship','Handcap','N
```

In [6]:
```
#Change "Yes" and "No" in No-Show to "0" and "1" to be consistent with other values tha
brazilmed_df.replace(to_replace =["No","Yes"], value =["0","1"], inplace=True)
brazilmed_df['No-show'] = brazilmed_df['No-show'].astype(int64)
```

In [7]:
```
#Rename No-show to No_show to avoid possible syntax errors
brazilmed_df.rename(columns={'No-show': 'No_show'}, inplace=True)
```

In [8]:
```
#Confirm that only data relevant to this project is remaining and the values are the ty
brazilmed_df.info()
brazilmed_df.head()
```

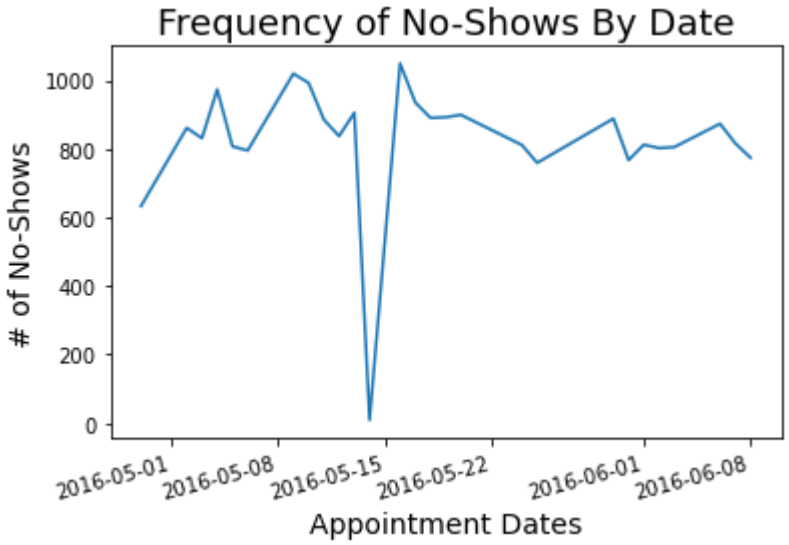
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   AppointmentID   110527 non-null  int64
 1   AppointmentDay  110527 non-null  datetime64[ns, UTC]
 2   Hipertension    110527 non-null  int64
 3   Diabetes        110527 non-null  int64
 4   Alcoholism      110527 non-null  int64
 5   SMS_received    110527 non-null  int64
 6   No_show         110527 non-null  int64
dtypes: datetime64[ns, UTC](1), int64(6)
memory usage: 5.9 MB
```

Out[8]:

	AppointmentID	AppointmentDay	Hipertension	Diabetes	Alcoholism	SMS_received	No_show
0	5642903	2016-04-29 00:00:00+00:00	1	0	0	0	0
1	5642503	2016-04-29 00:00:00+00:00	0	0	0	0	0
2	5642549	2016-04-29 00:00:00+00:00	0	0	0	0	0
3	5642828	2016-04-29 00:00:00+00:00	0	0	0	0	0
4	5642494	2016-04-29 00:00:00+00:00	1	1	0	0	0

Exploratory Data Analysis

In [9]:
```
#Observe no-shows across a period of time
noshows_df=brazilmed_df.groupby('AppointmentDay')['No_show'].sum()
noshows_df.plot.line()
plt.xticks(rotation=15)
plt.xlabel('Appointment Dates',fontsize=14)
plt.ylabel('# of No-Shows',fontsize=14)
plt.title('Frequency of No-Shows By Date',fontsize=18);
```



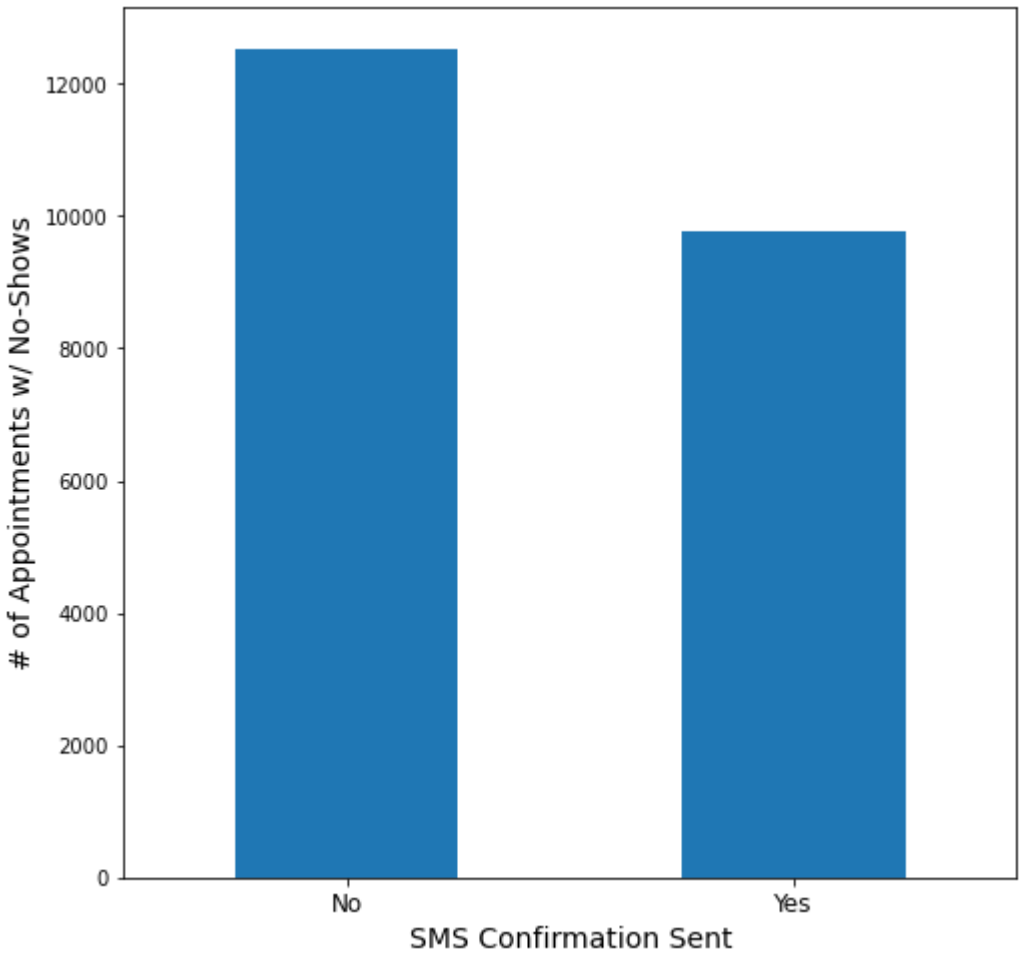
Displayed in the line chart above are all no-shows by appointment day.

Does SMS confirmation reduce no-shows?

In [10]:
```
#Determine how many of the no-shows were sent text messages to confirm their appointmen
brazilmed_df.groupby('SMS_received')['No_show'].sum()
```

Out[10]:
```
SMS_received
0    12535
1     9784
Name: No_show, dtype: int64
```

In [11]:
```
#Chart the totals to see it represented visually
brazilmed_df.groupby('SMS_received')['No_show'].sum().plot(kind='bar', rot=0, figsize=(
plt.xlabel("SMS Confirmation Sent",fontsize=14)
plt.ylabel("# of Appointments w/ No-Shows",fontsize=14);
```



By comparing appointment no-shows sent SMS confirmations to the ones that weren't and then plotting it on a bar chart, it was apparent that there were fewer no-shows for appointments where an SMS confirmation was sent.

How many of the no-shows are appointments for patients with non-handicap medical conditions?

In [12]:
```
#Get the total number of no-shows as a baseline
noshows_df=brazilmed_df[brazilmed_df['No_show'] ==1]
nstotal=noshows_df['No_show'].value_counts()
nstotal
```

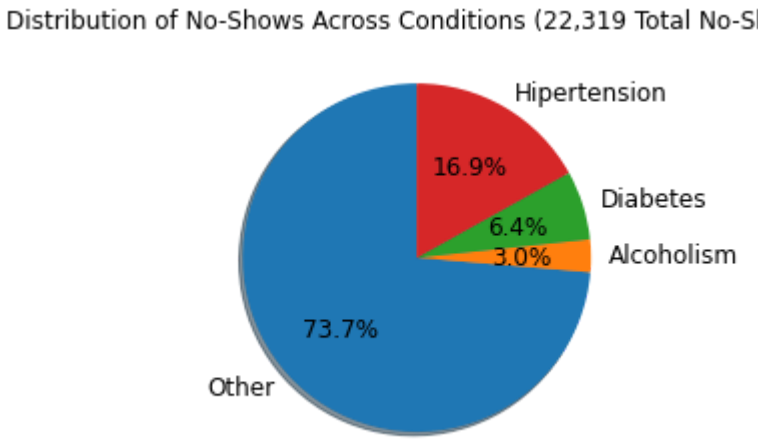
Out[12]:
```
1    22319
Name: No_show, dtype: int64
```

In [13]:
```
#Create a function to find out how many of the no-shows have a medical condition
def condns(Condition):
    condnoshow=noshows_df[noshows_df[(Condition)] == 1]
    nscondtotal= condnoshow['No_show'].value_counts()
    return (nscondtotal)
```

In [14]:
```
#Subtract the total of no-showed appointments involving patients with listed medical co
othernsreasonstotal = nstotal - (condns('Alcoholism') + condns('Diabetes') + condns('Hi
```

In [15]:
```
#Use a pie chart to show the total of each category as a percentage of the entire no-sh
df_medc_pie = pd.DataFrame( {'Totals': [othernsreasonstotal, condns('Alcoholism'), cond
```

```
df_medc_pie.astype(int64).plot(kind='pie', y=0, autopct='%1.1f%%', shadow=True, startan
plt.title('Distribution of No-Shows Across Conditions (22,319 Total No-Shows)');
```



Using a three variable calculation (unique appointment identifiers, non-handicap medical condition, and whether or not the patient no-showed), I was able to determine the distribution and plot it as a pie-chart.

26.3% of no-shows were for appointments where the patient had a non-handicap medical condition.

Are the no-shows for alcoholism proportionate to their percentage of all appointments?

In [27]:
```
#Find the total of all appointments where the patient is being treated for alcoholism.
#Divide it by the total number of all appointments to get a percentage
alctotal_df=brazilmed_df[brazilmed_df['Alcoholism']==1]
alctotal=alctotal_df.value_counts().sum()
allappts=brazilmed_df['AppointmentID'].value_counts().sum()
alcpropofall=round(alctotal/allappts,2)
alcpropofallpct="{:.0%}".format(alcpropofall)
alcpropofallpct
```

Out[27]: '3%'

In [17]:
```
#Subtract appointments for patients being treated for alcoholism from the total of all
alctotal=alctotal_df.value_counts().sum()
allotherappts = allappts - alctotal
```

In [18]:
```
#Plot the results on a pie chart to compare how many alcoholism appointments make up th
df_medalc_pie = pd.DataFrame( {'Totals': [alctotal,allotherappts] }, index=['Alcoholism
df_medalc_pie.astype(int64).plot(kind='pie', y=0, autopct='%1.1f%%', shadow=True, start
plt.title('Proportion of Alcoholism Treatment Appointments');
```

In order to determine if alcoholism contributed to more no-shows, all no-shows with a patient being treated for alcoholism were compared with all appointments for a patient being treated for alcoholism. This is displayed in the pie chart above.

No-shows for appointments where patients were being treated for alcoholism made up approximately 3% of the entire no-show total.
All appointments for patients being treated for alcoholism made up approximately 3% of all appointments in the dataset.
Based on the visualizations above, substance use as defined by this dataset does not seem to have a disproportionate affect on no-shows.

Conclusions

For appointments where an SMS confirmation was sent beforehand, there were fewer no-shows.

Appointments where the patient did not have a medical condition, regardless of handicap, made up almost 3/4 of all no-shows. This was somewhat expected as people not needing ongoing treatment for an established medical condition that requires monitoring would presumably be more likely to no-show.

Substance use did not appear to disproportionately contribute to no-shows. The distributions were roughly the same.

Limitations

Researching whether or not SMS confirmation impacts a no-show was only able to be done on appointments where it was attempted. Not all appointments implemented this service.

Determining the proportion of how many no-show appointments were from people with at least one non-handicap medical condition did not take into account those who had one or more.

Where a patient was at in their alcoholism treatment is unknown. This could impact whether or not they were just as likely to no-show as non-alcoholics.