For my project I analysed the noshowappointments-kagglev2-may-2016 dataset

The questions posed during the data investigation include;
1. Does gender influence the rate of attendance?
2. Does age determine probability of attendance?
3. Do those people who receive SMS attend regularly?
4. Does any medical condition accelerate probability of attendance?

For question 1 (Does gender influence the rate of attendance

Since I was interested in investigating the correlation between a specific variable and the final outcome (show vs no-show), it was worth writing a function that will be used multiple times.

The function will cross reference the desired variable against the final outcome (column Attended), normalizing the result between Yes-Attended and No-Attended and plotting a bar chart with the respective title and labels:

```
def proportion_attendance(feature):
    pd.crosstab(df[feature], df.attended, normalize='columns').plot(kind='bar', alpha=0.85)
    plt.xlabel('Feature: {}'.format(feature.replace('_',' ').title()))
    plt.ylabel('Proportion')
    plt.title('Proportion of Attendance Rate by {}'.format(feature.replace('_',' ').title()))
    plt.margins(y=0.1)
    return plt # this allows  to plot  a bar graph which is used to  check  the correlation
```

To answer this question I used the below code to plot the bar graph

```
proportion_attendance('gender');
```

A bar graph with the X Axis labelled as Gender and the y axis labelled as proportion.
From the graph I could deduce whether there is any correlation between gender and rate of attendance.

For question 2 Does age determine probability of attendance?
I was also able to plot a bar graph so as to theck whether there is a correlation between age and the rate of attendance.
Using the function defined previously. I used the below code for plotting the graphA and since the age distribution is so spread out, let's transform this numeric variable into age groups - representing tenth percentile each - to further investigate this question:

```
df['age_groups'] = pd.qcut(df.age, 10)
proportion_attendance('age_groups');
```

For question 3 Do those people who receive SMS attend regularly?
I also used the previously declared function and the below code to plot a bar graph with the x axis representing sms received and the y axis representing the proportion of attendance.

```
proportion_attendance('sms_received');
```

For question 4 Does any medical condition accelerate probability of attendance?
I also used the previously declared function and the below code to plot a bar graph

with the x axis representing the condition and the y axis representing the proportion of attendance.

```
proportion_attendance('alcoholism');
proportion_attendance('handcap');
proportion_attendance('diabetes');
proportion_attendance('hipertension');
```

From the above bar graphs I could be able to tell which condition leads to a higher rate of attendance.

Data wrangling

First I was able to import or read the data using the code below

```
df = pd.read_csv("Dataset - NoShowAppointments.csv")
df.head(3)
```

I was able to Change the column titles to be in uppercase so as to be clearly visible

```
df.columns.str.upper()
```

checking for any missing values and datatypes apparently there are none using the below code

```
df.info()
```

I also had to check for any duplicated values using the below code

```
df.duplicated().sum() , df.duplicated('patient_id').sum(),
df.duplicated('appointment_id').sum()
```

The columns "Patient ID" and "Appointment ID" doesn't seem very promising to our analysis, so I removed them :

```
df.drop(['patient_id','appointment_id'],axis=1, inplace=True)
```

Looking at the overall description of the data, the column Age seems off: the minimum value is -1, which doesn't make sense. In addition, the maximum value is 115, which seems too high. We need to investigate for possible outlier and remove negative values:

```
df[df.age < 0]
# Box-plot to visualize threshold for possible outliers
df.age.plot(kind='box');
iqr = df.age.quantile(0.75) - df.age.quantile(0.25)
outlier = df.age.quantile(0.75) + 1.5*iqr
print('Outlier Threshold:', int(round(outlier,0)))
df[df.age > 110]
```

I also tweaked this dataset a little more by renaming the column "no_show" as "attended", since it is more natural to think of 'yes' as a positive observation. using the below code

```
df.rename(columns={'no_show': 'PRESENT'}, inplace=True)
df.attended.replace(['No', 'Yes'], [1, 0], inplace=True)
df.head(1)
```

Conclusions and limitations

After analyzing the data and answering the investigation questions we can conclude that some features impact the probability of patient’s attendance. Some features

showed a high attendance rate compared to others. This descriptive statistical analysis allows us to identify features with possible correlation to the dependent variable - which would be interesting to further investigate using hypothesis testing and/or regression models.

Here is a summary of the initial findings:

- Gender does not seem to have any impact with the attendance rate;
- Age groups:
  - Less than 5 years and over 45 years old: seems to be more likely to show up
  - Between 5 and 45 years old: seems to be less likely to show up
- Patients receiving Scholarship (Bolsa Familia) are less likely to show up;
- Medical Conditions: patients suffering from alcoholism and diabetes seems to be slightly more likely to attend
- SMS reminder: even though there is evidence receiving an SMS reminder seems to make the patients less likely to show up, we should ignore this variable due to lack of explain ability.

Limitations

Missing features that could be useful to get surer what is the most feature that impacts showing to the appointment such as if the patient is employed or not, or whether the patient have a series medical issue or not. There we some illogical data such as patients with age 0 or less