Project: Investigate a Medical Dataset

Table of Contents

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
% matplotlib inline
import seaborn as sns
```

Introduction

Why do 20%+ of patients miss their scheduled appointments? This project analyzes data collected from 100k medical appointments in Brazil and is focused on the question of whether or not patients show up for their appointment. Using NumPy and Pandas, this project looks at relationships between multiple variables: at least one dependent variable (No-show) and three independent variables (age, diabetes, week_day). variables are investigated using both single-variable (1d) and multiple-variable (2d) explorations.

Data:

- 'ScheduledDay' tells us on what day the patient set up their appointment.
- 'Neighbourhood' indicates the location of the hospital.
- 'Scholarship' indicates whether or not the patient is enrolled in the Brasilian scholarship program.
- 'No-show' says 'No' if the patient showed up to their appointment, and 'Yes' if they did not show up.

Questions

- What is the overall no-show percentage?
- Which factors can help predict if a patient will miss their scheduled appointment?
- What is the relationship between absenteeism and age?
- What is the relationship between absenteeism and appointment day?

Data wrangling

General properties

In [2]:

```
# Load data and print out a few lines. Perform operations to inspect data
#   types and look for instances of missing or possibly errant data.
df = pd.read_csv('noshowappointments-kagglev2-may-2016.csv')
df.head()
```

Out[2]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension
[illegible]	[illegible]	[illegible]	[illegible]	2016-04-	2016-04-	[illegible]	JARDIM DA	[illegible]	[illegible]

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

In [3]:

```
df.shape
```

Out[3]:

```
(110527, 14)
```

In [4]:

```
print("Rows:  ", df.shape[0])
print("Columns: ", df.shape[1])
```

```
Rows:   110527
Columns:  14
```

In [5]:

```
#Review data types
df.dtypes
```

Out[5]:

```
PatientId         float64
AppointmentID       int64
Gender             object
ScheduledDay       object
AppointmentDay     object
Age                 int64
Neighbourhood      object
Scholarship         int64
Hipertension        int64
Diabetes            int64
Alcoholism          int64
Handcap             int64
SMS_received        int64
No-show            object
dtype: object
```

In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
PatientId         110527 non-null float64
AppointmentID     110527 non-null int64
Gender            110527 non-null object
ScheduledDay      110527 non-null object
AppointmentDay    110527 non-null object
Age               110527 non-null int64
Neighbourhood     110527 non-null object
Scholarship       110527 non-null int64
Hipertension      110527 non-null int64
Diabetes          110527 non-null int64
Alcoholism        110527 non-null int64
Handcap           110527 non-null int64
```

```
SMS_received      110527 non-null int64
No-show           110527 non-null object
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
```

In [7]:

```
#See entry types and options for each column
for  feature  in  ["Diabetes", "Alcoholism", "Hipertension", "Handcap",
                "Scholarship", "SMS_received", "Neighbourhood", 'No-show']:
    print("{}: {}".format(feature, df[feature].unique()))
```

```
Diabetes: [0 1]
Alcoholism: [0 1]
Hipertension: [1 0]
Handcap: [0 1 2 3 4]
Scholarship: [0 1]
SMS_received: [0 1]
Neighbourhood: ['JARDIM DA PENHA' 'MATA DA PRAIA' 'PONTAL DE CAMBURI' 'REPÚBLICA'
 'GOIABEIRAS' 'ANDORINHAS' 'CONQUISTA' 'NOVA PALESTINA' 'DA PENHA'
 'TABUAZEIRO' 'BENTO FERREIRA' 'SÃO PEDRO' 'SANTA MARTHA' 'SÃO CRISTÓVÃO'
 'MARUÍPE' 'GRANDE VITÓRIA' 'SÃO BENEDITO' 'ILHA DAS CAIEIRAS'
 'SANTO ANDRÉ' 'SOLON BORGES' 'BONFIM' 'JARDIM CAMBURI' 'MARIA ORTIZ'
 'JABOUR' 'ANTÔNIO HONÓRIO' 'RESISTÊNCIA' 'ILHA DE SANTA MARIA'
 'JUCUTUQUARA' 'MONTE BELO' 'MÁRIO CYPRESTE' 'SANTO ANTÔNIO' 'BELA VISTA'
 'PRAIA DO SUÁ' 'SANTA HELENA' 'ITARARÉ' 'INHANGUETÁ' 'UNIVERSITÁRIO'
 'SÃO JOSÉ' 'REDENÇÃO' 'SANTA CLARA' 'CENTRO' 'PARQUE MOSCOSO' 'DO MOSCOSO'
 'SANTOS DUMONT' 'CARATOÍRA' 'ARIOVALDO FAVALESSA' 'ILHA DO FRADE'
 'GURIGICA' 'JOANA D´ARC' 'CONSOLAÇÃO' 'PRAIA DO CANTO' 'BOA VISTA'
 'MORADA DE CAMBURI' 'SANTA LUÍZA' 'SANTA LÚCIA' 'BARRO VERMELHO'
 'ESTRELINHA' 'FORTE SÃO JOÃO' 'FONTE GRANDE' 'ENSEADA DO SUÁ'
 'SANTOS REIS' 'PIEDADE' 'JESUS DE NAZARETH' 'SANTA TEREZA' 'CRUZAMENTO'
 'ILHA DO PRÍNCIPE' 'ROMÃO' 'COMDUSA' 'SANTA CECÍLIA' 'VILA RUBIM'
 'DE LOURDES' 'DO QUADRO' 'DO CABRAL' 'HORTO' 'SEGURANÇA DO LAR'
 'ILHA DO BOI' 'FRADINHOS' 'NAZARETH' 'AEROPORTO'
 'ILHAS OCEÂNICAS DE TRINDADE' 'PARQUE INDUSTRIAL']
No-show: ['No' 'Yes']
```

In [8]:

```
#Check to see if there is any missing data
df.isnull().any()
```

Out[8]:

```
PatientId         False
AppointmentID     False
Gender            False
ScheduledDay      False
AppointmentDay    False
Age               False
Neighbourhood     False
Scholarship       False
Hipertension      False
Diabetes          False
Alcoholism        False
Handcap           False
SMS_received      False
No-show           False
dtype: bool
```

Check for duplicates and review uniques

In [9]:

```
df.duplicated().sum()
```

Out[9]:

```
0
```

In [10]:

```
df.nunique().sum()
```

Out[10]:

```
276606
```

Data cleaning

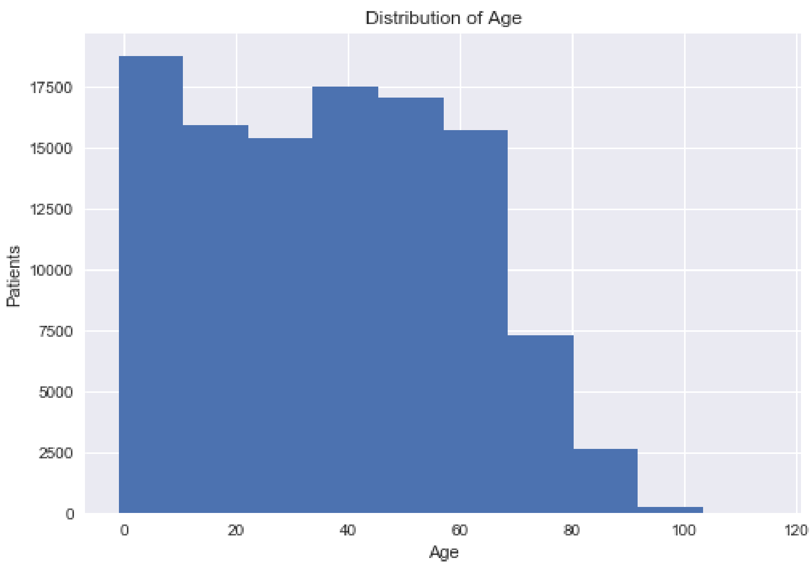
In [11]:

```
# Checking 'Age' entries for outliers
print ("Age range:", sorted(df['Age'].unique()))
```

```
Age range: [-1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22,
23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 4
7, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71
, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95,
96, 97, 98, 99, 100, 102, 115]
```

In [12]:

```
# Review 'Age' distribution
plt.figure();
age_hist = df['Age'].plot.hist(bins=10)
age_hist.set_xlabel("Age")
age_hist.set_ylabel("Patients")
age_hist.set_title('Distribution of Age');
```



In [13]:

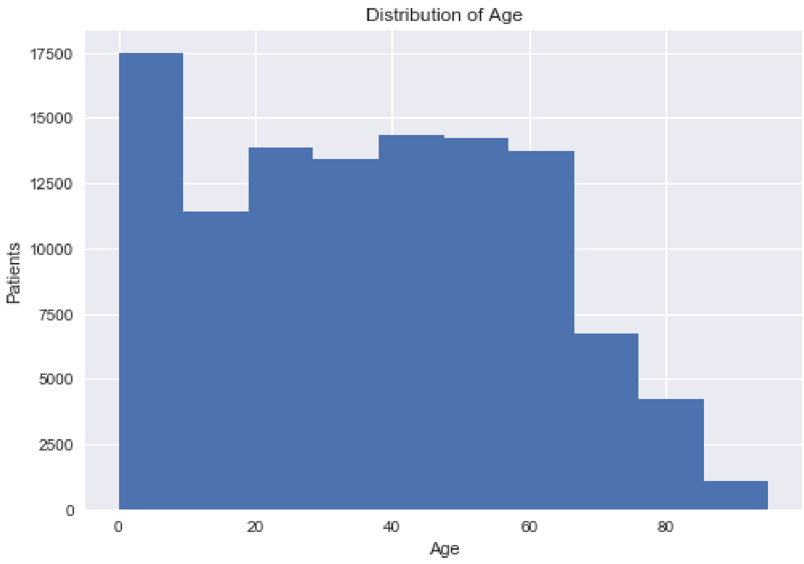
```
# remove age outliers
df = df[(df.Age >= 0) & (df.Age <= 95)]
df.shape
```

Out[13]:

```
(110480, 14)
```

In [14]:

```
# Distribution of 'Age'
plt.figure();
age_hist = df['Age'].plot.hist(bins=10)
age_hist.set_xlabel("Age")
age_hist.set_ylabel("Patients")
age_hist.set_title('Distribution of Age');
```



In [15]:

```
min_age = df['Age'].min()
max_age = df['Age'].max()
print ("Age now spans: {} to {}.".format(min_age, max_age))
```

```
Age now spans: 0 to 95.
```

In [16]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 110480 entries, 0 to 110526
Data columns (total 14 columns):
PatientId         110480 non-null float64
AppointmentID     110480 non-null int64
Gender            110480 non-null object
ScheduledDay      110480 non-null object
AppointmentDay    110480 non-null object
Age               110480 non-null int64
Neighbourhood     110480 non-null object
Scholarship       110480 non-null int64
Hipertension      110480 non-null int64
Diabetes          110480 non-null int64
Alcoholism        110480 non-null int64
Handcap           110480 non-null int64
SMS_received      110480 non-null int64
No-show           110480 non-null object
dtypes: float64(1), int64(8), object(5)
memory usage: 12.6+ MB
```

In [17]:

```
#Make all column headers lower case
df.columns = [x.lower() for x in df.columns]
```

In [18]:

```
#Drop 'scheduleday' as we will not need it to investigate
#the specificed questions.
df = df.drop('scheduledday', 1)
```

In [19]:

```
#Replace dashes with underscores
df.columns = [x.strip().replace('-', '_') for x in df.columns]
```

In [20]:

```
# Renaming columns to fix typos, inconsistency and spelling mistakes
df.rename(columns={'appointmentid' :'appointment_id',
                   'handcap' :'handicap',
                   'patientid' : 'patient_id',
                   'appointmentid' : 'appointment_id',
                   'appointmentday' : 'appointment_day',
                   'neighbourhood' : 'neighborhood',
                   'no_show' : 'absenteeism',
                   'hipertension':'hypertension'},
          inplace=True
         )
```

In [21]:

```
df.head()
```

Out[21]:

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes	alcoh
0	2.987250e+13	5642903	F	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0
1	5.589978e+14	5642503	M	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0
2	4.262962e+12	5642549	F	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0
3	8.679512e+11	5642828	F	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0
4	8.841186e+12	5642494	F	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0

Manage absenteeism column

In [22]:

```
#Replace 'Yes' and 'No' with 1 and 0 for 'No-show'
  #0 = Showed up to appointment
  #1 = did not show up to appointment (missed it)
```

In [23]:

```
#df['absenteeism'] = df['absenteeism'].map({'Yes':1,'No':0})
```

In [24]:

```
df['absenteeism'].replace({'No':0,'Yes':1},inplace=True)
```

In [25]:

```
# Create variables for missed and arrived
arrived = df.absenteeism == 0
missed = df.absenteeism == 1
```

Add a mask

In [26]:

```
df_attendance = df [df['absenteeism']==0]
```

In [27]:

In [27]:

```
mask = df['absenteeism'] == 0
df_attendance = df[mask]
df_attendance
```

Out[27]:

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
0	2.987250e+13	5642903	F	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0
1	5.589978e+14	5642503	M	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0
2	4.262962e+12	5642549	F	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0
3	8.679512e+11	5642828	F	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0
4	8.841186e+12	5642494	F	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1
5	9.598513e+13	5626772	F	2016-04-29T00:00:00Z	76	REPÚBLICA	0	1	0
8	5.639473e+13	5638447	F	2016-04-29T00:00:00Z	21	ANDORINHAS	0	0	0
9	7.812456e+13	5629123	F	2016-04-29T00:00:00Z	19	CONQUISTA	0	0	0
10	7.345362e+14	5630213	F	2016-04-29T00:00:00Z	30	NOVA PALESTINA	0	0	0
12	5.666548e+14	5634718	F	2016-04-29T00:00:00Z	22	NOVA PALESTINA	1	0	0
13	9.113946e+14	5636249	M	2016-04-29T00:00:00Z	28	NOVA PALESTINA	0	0	0
14	9.988472e+13	5633951	F	2016-04-29T00:00:00Z	54	NOVA PALESTINA	0	0	0
15	9.994839e+10	5620206	F	2016-04-29T00:00:00Z	15	NOVA PALESTINA	0	0	0
16	8.457439e+13	5633121	M	2016-04-29T00:00:00Z	50	NOVA PALESTINA	0	0	0
18	1.713538e+13	5621836	F	2016-04-29T00:00:00Z	30	NOVA PALESTINA	1	0	0
19	7.223289e+12	5640433	F	2016-04-29T00:00:00Z	46	DA PENHA	0	0	0
23	2.137540e+14	5634142	F	2016-04-29T00:00:00Z	46	CONQUISTA	0	0	0
24	8.734858e+12	5641780	F	2016-04-29T00:00:00Z	65	TABUAZEIRO	0	0	0
25	5.819370e+12	5624020	M	2016-04-29T00:00:00Z	46	CONQUISTA	0	1	0
26	2.578785e+10	5641781	F	2016-04-29T00:00:00Z	45	BENTO FERREIRA	0	1	0
27	1.215484e+13	5628345	F	2016-04-29T00:00:00Z	4	CONQUISTA	0	0	0
28	5.926172e+12	5642400	M	2016-04-29T00:00:00Z	51	SÃO PEDRO	0	0	0
29	1.225776e+12	5642186	F	2016-04-29T00:00:00Z	32	SANTA MARTHA	0	0	0
				2016-04-		NOVA			

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
30	3.428156e+11	5628068	F	2016-04-29T00:00:00Z	46	NOVA PALESTINA	0	0	0
32	5.288356e+13	5637908	M	2016-04-29T00:00:00Z	61	SÃO CRISTÓVÃO	0	1	0
33	7.653517e+12	5616921	F	2016-04-29T00:00:00Z	38	SÃO CRISTÓVÃO	1	0	0
34	1.999976e+13	5637963	F	2016-04-29T00:00:00Z	79	SÃO CRISTÓVÃO	0	1	0
35	7.816264e+13	5637968	M	2016-04-29T00:00:00Z	18	SÃO CRISTÓVÃO	0	0	0
36	7.298459e+13	5637975	F	2016-04-29T00:00:00Z	63	SÃO CRISTÓVÃO	0	1	1
37	1.578132e+12	5637986	F	2016-04-29T00:00:00Z	64	TABUAZEIRO	1	1	1
...	...	...	...	...	...	...	...	...	...
110494	2.895817e+14	5779073	F	2016-06-08T00:00:00Z	38	MARIA ORTIZ	0	0	0
110495	7.499489e+12	5759838	M	2016-06-01T00:00:00Z	40	MARIA ORTIZ	0	0	0
110497	7.935892e+14	5757745	M	2016-06-01T00:00:00Z	76	MARIA ORTIZ	0	0	0
110498	9.433654e+13	5787655	F	2016-06-08T00:00:00Z	59	MARIA ORTIZ	0	0	0
110499	8.219692e+14	5757697	F	2016-06-01T00:00:00Z	66	MARIA ORTIZ	0	1	1
110500	4.434384e+14	5787233	F	2016-06-08T00:00:00Z	59	MARIA ORTIZ	0	0	0
110501	4.544252e+11	5758133	M	2016-06-01T00:00:00Z	44	MARIA ORTIZ	0	0	0
110502	7.316229e+14	5787937	F	2016-06-08T00:00:00Z	22	GOIABEIRAS	0	0	0
110503	2.362182e+13	5759473	F	2016-06-01T00:00:00Z	64	SOLON BORGES	0	0	0
110504	9.947983e+12	5788052	F	2016-06-08T00:00:00Z	4	MARIA ORTIZ	0	0	0
110505	5.667344e+13	5758455	F	2016-06-01T00:00:00Z	55	MARIA ORTIZ	0	0	0
110506	8.973883e+11	5758779	M	2016-06-01T00:00:00Z	5	MARIA ORTIZ	0	0	0
110507	4.769462e+14	5786918	F	2016-06-08T00:00:00Z	0	MARIA ORTIZ	0	0	0
110508	9.433654e+13	5757656	F	2016-06-01T00:00:00Z	59	MARIA ORTIZ	0	0	0
110509	4.952968e+14	5786750	M	2016-06-08T00:00:00Z	33	MARIA ORTIZ	0	0	0
110510	2.362182e+13	5757587	F	2016-06-01T00:00:00Z	64	SOLON BORGES	0	0	0
110511	8.235996e+11	5786742	F	2016-06-08T00:00:00Z	14	MARIA ORTIZ	0	0	0
110512	9.876246e+13	5786368	F	2016-06-08T00:00:00Z	41	MARIA ORTIZ	0	0	0
110513	8.674778e+13	5785964	M	2016-06-08T00:00:00Z	2	ANTÔNIO HONÓRIO	0	0	0
				2016-06-					

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
110514	2.695685e+12	5786567	F	2016-06-07T00:00:00Z	58	MARIA ORTIZ	0	0	0
110517	5.574942e+12	5780122	F	2016-06-07T00:00:00Z	19	MARIA ORTIZ	0	0	0
110518	7.263315e+13	5630375	F	2016-06-07T00:00:00Z	50	MARIA ORTIZ	0	0	0
110519	6.542388e+13	5630447	F	2016-06-07T00:00:00Z	22	MARIA ORTIZ	0	0	0
110520	9.969977e+14	5650534	F	2016-06-07T00:00:00Z	42	MARIA ORTIZ	0	0	0
110521	3.635534e+13	5651072	F	2016-06-07T00:00:00Z	53	MARIA ORTIZ	0	0	0
110522	2.572134e+12	5651768	F	2016-06-07T00:00:00Z	56	MARIA ORTIZ	0	0	0
110523	3.596266e+12	5650093	F	2016-06-07T00:00:00Z	51	MARIA ORTIZ	0	0	0
110524	1.557663e+13	5630692	F	2016-06-07T00:00:00Z	21	MARIA ORTIZ	0	0	0
110525	9.213493e+13	5630323	F	2016-06-07T00:00:00Z	38	MARIA ORTIZ	0	0	0
110526	3.775115e+14	5629448	F	2016-06-07T00:00:00Z	54	MARIA ORTIZ	0	0	0

88168 rows × 13 columns

In [28]:

```
df_attendance.head()
```

Out[28]:

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes	alcoh
0	2.987250e+13	5642903	F	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0
1	5.589978e+14	5642503	M	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0
2	4.262962e+12	5642549	F	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0
3	8.679512e+11	5642828	F	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0
4	8.841186e+12	5642494	F	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0

Manage gender entries

In [29]:

```
#Replace 'M' and 'F' with 1 and 0 for 'Gender'
#df['gender'] = df['gender'].map({'M':1,'F':0})
```

In [30]:

```
df['gender'].replace({'M':0,'F':1},inplace=True)
```

convert to datetime; add: weekday, total_missed

In [31]:

```
df.appointment_day = df.appointment_day.apply(np.datetime64)
```

In [32]:

```
df['week_day'] = pd.to_datetime(df['appointment_day']).apply(lambda x: x.isoweekday())
```

Adding age bins

In [33]:

```
#Creating the age_bins
bins = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
df['age_bins'] = pd.cut(df['age'], bins)
```

Exploratory analysis

Descriptive summary

In [34]:

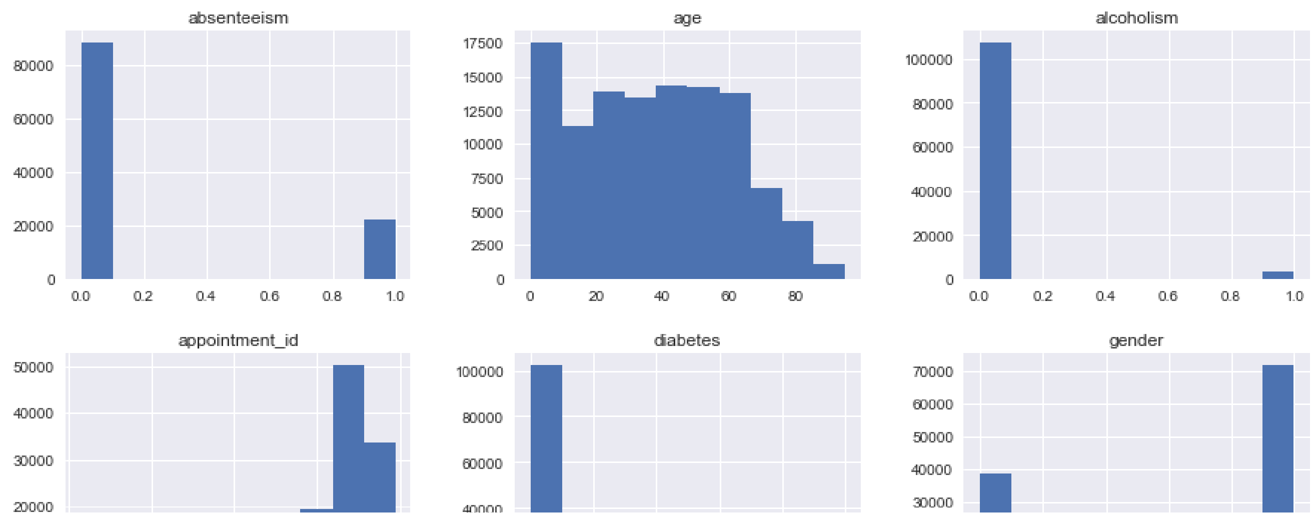
```
df.describe()
```

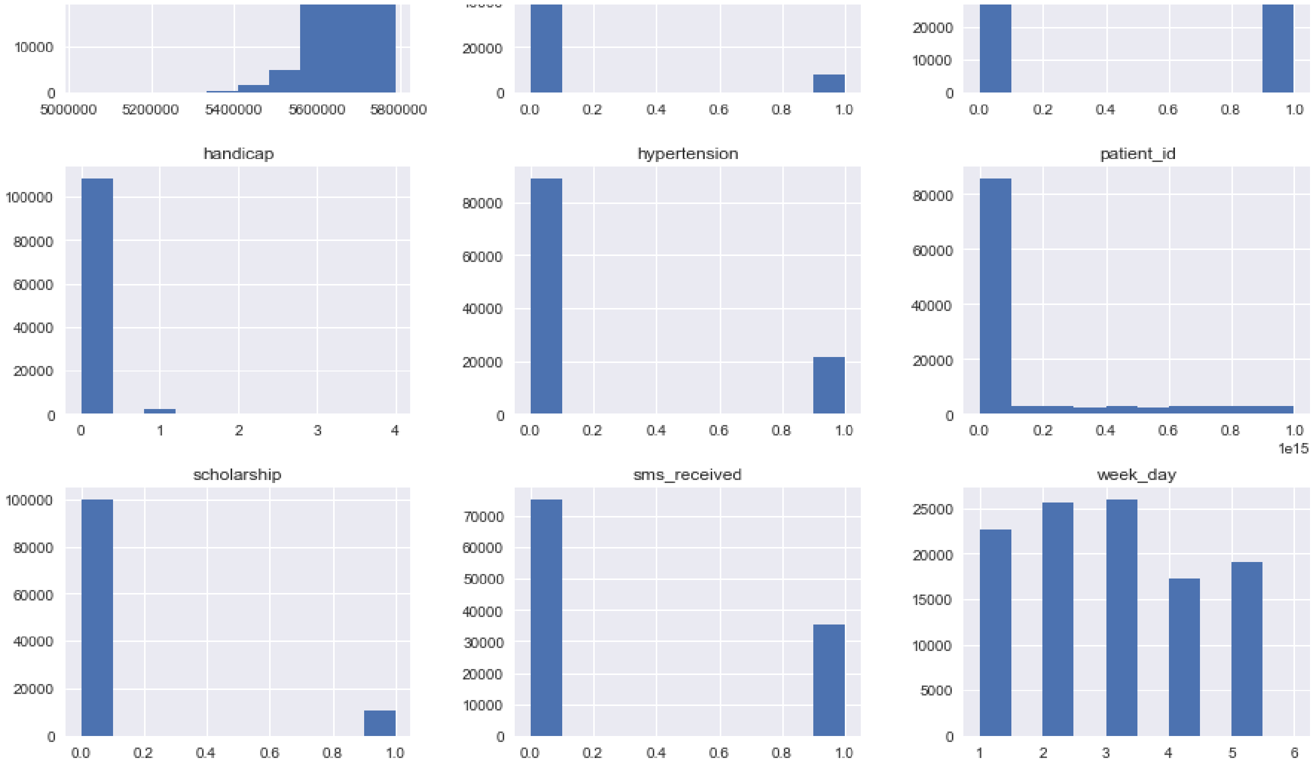
Out[34]:

	patient_id	appointment_id	gender	age	scholarship	hypertension	diabetes	al
count	1.104800e+05	1.104800e+05	110480.000000	110480.000000	110480.000000	110480.000000	110480.000000	11048
mean	1.474691e+14	5.675303e+06	0.649909	37.063342	0.098307	0.197076	0.071841	0.0304
std	2.560626e+14	7.128285e+04	0.477000	23.079712	0.297731	0.397792	0.258226	0.1717
min	3.921784e+04	5.030230e+06	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
25%	4.172457e+12	5.640284e+06	0.000000	18.000000	0.000000	0.000000	0.000000	0.0000
50%	3.172598e+13	5.680564e+06	1.000000	37.000000	0.000000	0.000000	0.000000	0.0000
75%	9.438179e+13	5.725507e+06	1.000000	55.000000	0.000000	0.000000	0.000000	0.0000
max	9.999816e+14	5.790484e+06	1.000000	95.000000	1.000000	1.000000	1.000000	1.0000

In [35]:

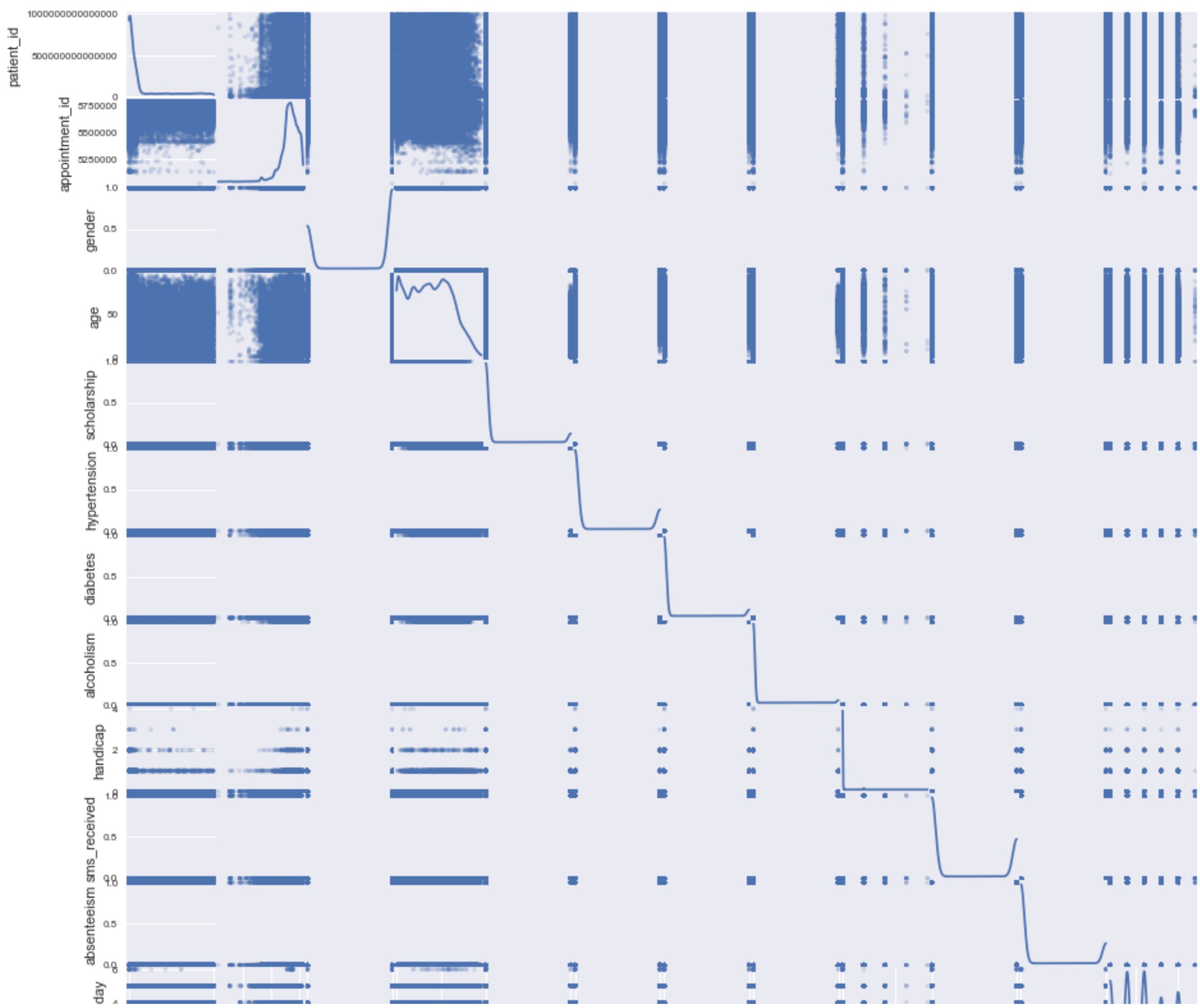
```
#Review and discover the underlying frequency distribution (shape)
#of the data points in each column
df.hist(figsize=(15,15));
```

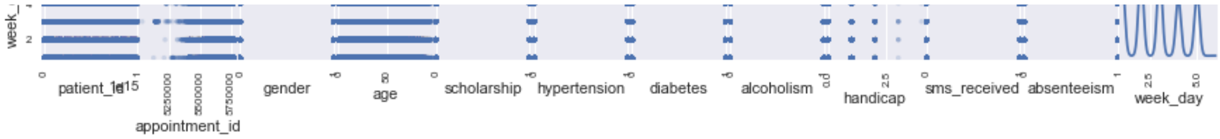




In [36]:

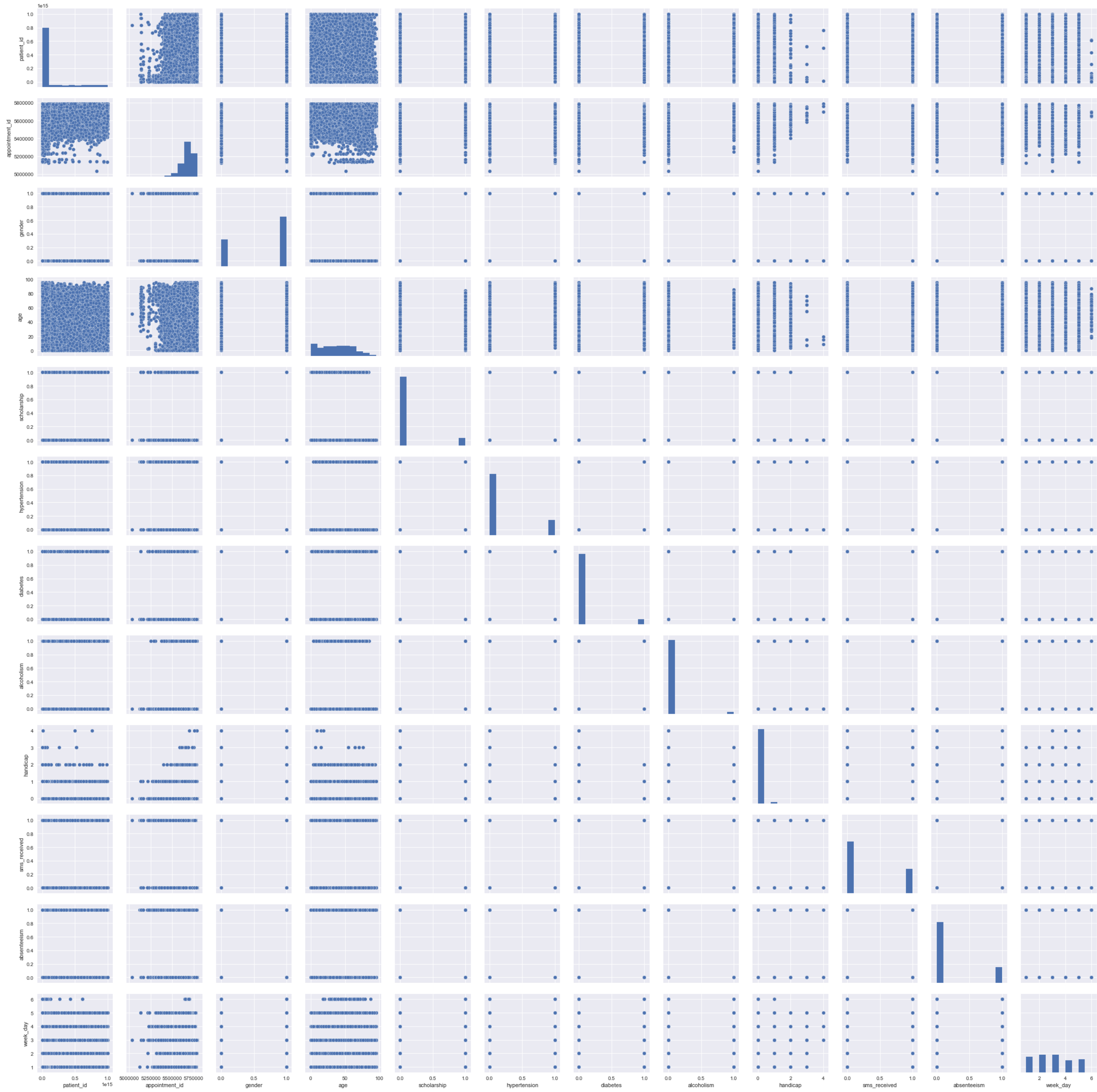
```python
# Review the relationships between the attributes
from pandas.plotting import scatter_matrix
scatter_matrix(df, alpha=0.2, figsize=(15, 15), diagonal='kde');
```





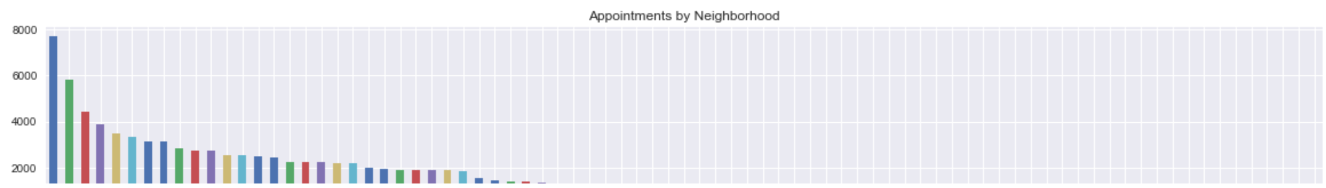
In [37]:

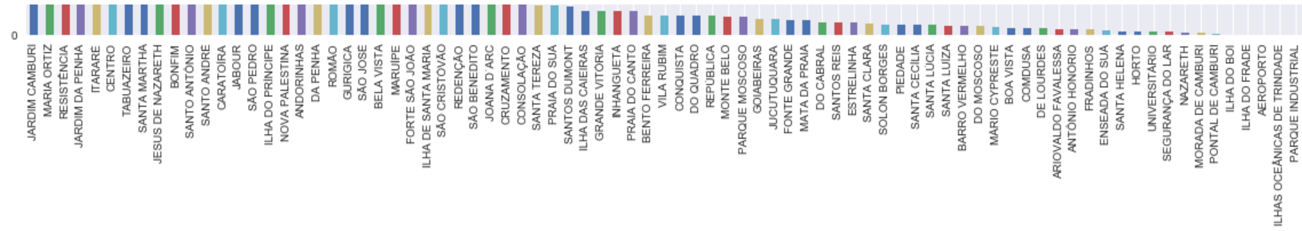
```
#Review the interactions between the variables
sns.pairplot(df);
```



In [38]:

```
#Discover which neighborhoods are most represented in the data set
df['neighborhood'].value_counts().plot(kind='bar', figsize=(20,3))
plt.title('Appointments by Neighborhood')
plt.show();
```





In [39]:

```
#create a list of top 30 neighborhoods to get a better view of those
#neighborhoods with highest number of appointments in the data set
top_30_areas_list = list(df['neighborhood'].value_counts().nlargest(30).index)
print(top_30_areas_list)
```

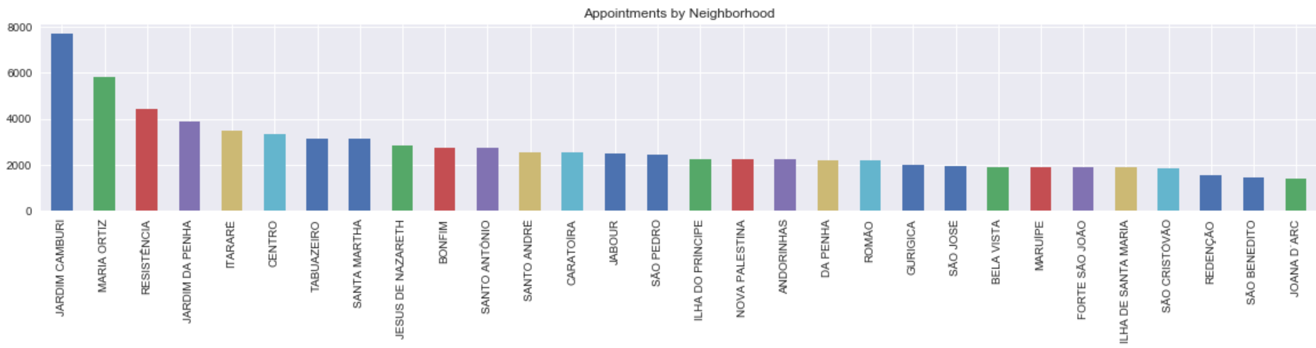
```
['JARDIM CAMBURI', 'MARIA ORTIZ', 'RESISTÊNCIA', 'JARDIM DA PENHA', 'ITARARÉ', 'CENTRO',
'TABUAZEIRO', 'SANTA MARTHA', 'JESUS DE NAZARETH', 'BONFIM', 'SANTO ANTÔNIO', 'SANTO ANDRÉ', 'CARA
TOÍRA', 'JABOUR', 'SÃO PEDRO', 'ILHA DO PRÍNCIPE', 'NOVA PALESTINA', 'ANDORINHAS', 'DA PENHA',
'ROMÃO', 'GURIGICA', 'SÃO JOSÉ', 'BELA VISTA', 'MARUÍPE', 'FORTE SÃO JOÃO', 'ILHA DE SANTA MARIA',
'SÃO CRISTÓVÃO', 'REDENÇÃO', 'SÃO BENEDITO', 'JOANA D´ARC']
```

In [40]:

```
#create dataframe for largest neighborhoods list
df_top_30_areas = df[df['neighborhood'].isin (top_30_areas_list)]
```

In [41]:

```
#visualize the top 30 neighborhoods by appointments
df_top_30_areas['neighborhood'].value_counts().plot(kind='bar', figsize=(20,3))
plt.title('Appointments by Neighborhood')
plt.show();
```



Questions

1. What is the overall no-show percentage?

> **Overall Absenteeism:** The overall no-show percentage is 20% for a total of 110480 records.

In [42]:

```
absent_total = df['absenteeism'].value_counts()
print(absent_total)

absent_percentage = absent_total[1]/ absent_total.sum() * 100
print("Percent who miss their appointments:", absent_percentage)
```

```
0    88168
1    22312
```

```
Name: absenteeism, dtype: int64
Percent who miss their appointments: 20.1955104996
```

2. Factors: Which factors can help predict if a patient will miss their scheduled appointment?

> **Most indicative factors:** The characteristics that show the highest correlation with missed appointments are sms_received and handicap(4).

Exploring absenteeism and attendance (arrived and missed)

In [43]:

```
columns = ['hypertension','alcoholism','diabetes', 'sms_received', 'handicap', 'scholarship']
for r in columns :
    print(df.groupby(r)['absenteeism'].mean())
```

```
hypertension
0    0.209048
1    0.173058
Name: absenteeism, dtype: float64
alcoholism
0    0.201970
1    0.201488
Name: absenteeism, dtype: float64
diabetes
0    0.203641
1    0.180169
Name: absenteeism, dtype: float64
sms_received
0    0.167036
1    0.275777
Name: absenteeism, dtype: float64
handicap
0    0.202387
1    0.178466
2    0.203297
3    0.230769
4    0.333333
Name: absenteeism, dtype: float64
scholarship
0    0.198095
1    0.237363
Name: absenteeism, dtype: float64
```

In [44]:

```
df.groupby('sms_received')['absenteeism'].mean()
```

Out[44]:

```
sms_received
0    0.167036
1    0.275777
Name: absenteeism, dtype: float64
```

In [45]:

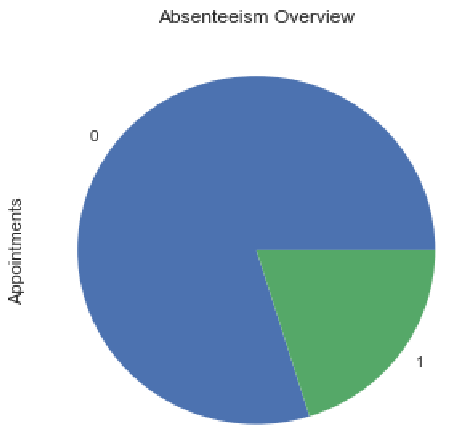
```
df['absenteeism'].value_counts()
```

Out[45]:

```
0    88168
1    22312
Name: absenteeism, dtype: int64
```

In [46]:

```
df['absenteeism'].value_counts().plot(kind='pie',figsize=(5,5));
plt.title('Absenteeism Overview')
plt.ylabel("Appointments");
```



Exploring absenteeism and hypertension, absenteeism and diabetes

In [47]:

```
#select all rows where the patient was absent and has hypertension
df[(df['hypertension'] == 1) & (df['absenteeism'] == 1)]
```

Out[47]:

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabete
44	2.741649e+11	5635414	1	2016-04-29	78	SÃO CRISTÓVÃO	0	1	1
212	4.266984e+14	5642059	0	2016-04-29	62	SANTOS DUMONT	0	1	1
270	8.617228e+12	5620528	1	2016-04-29	45	CARATOÍRA	1	1	0
272	5.119616e+12	5594085	1	2016-04-29	51	NOVA PALESTINA	0	1	0
352	5.917359e+12	5494413	1	2016-04-29	62	SÃO CRISTÓVÃO	0	1	0
441	3.935966e+13	5639484	0	2016-04-29	36	SÃO BENEDITO	0	1	0
463	5.228864e+08	5633872	1	2016-04-29	66	SÃO CRISTÓVÃO	0	1	1
537	3.546481e+13	5590085	0	2016-04-29	58	ITARARÉ	0	1	0
619	4.272659e+13	5634178	1	2016-04-29	67	PRAIA DO CANTO	0	1	1
684	8.238132e+13	5629179	1	2016-04-29	73	JOANA D´ARC	0	1	0
729	3.969538e+12	5636130	1	2016-04-29	69	CENTRO	0	1	1
751	6.468148e+14	5629130	0	2016-04-29	68	SÃO PEDRO	0	1	0
780	4.478798e+12	5611940	1	2016-04-29	69	JARDIM DA PENHA	0	1	0
914	1.899995e+13	5525204	0	2016-04-29	64	MARIA ORTIZ	0	1	0
951	7.414865e+12	5317449	1	2016-04-29	77	JESUS DE NAZARETH	0	1	0
968	7.244332e+14	5399572	1	2016-04-29	78	CRUZAMENTO	0	1	1
972	8.941338e+10	5361416	0	2016-04-29	75	BONFIM	0	1	1

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
973	8.21[illegible]e+14	[illegible]	1	2016-04-29	66	MARIA ORTIZ	0	1	1
974	1.729243e+14	5505270	0	2016-04-29	82	GRANDE VITÓRIA	0	1	0
975	7.584570e+14	5522365	0	2016-04-29	67	MARUÍPE	0	1	0
979	2.539578e+12	5627308	1	2016-04-29	81	SANTO ANDRÉ	0	1	1
1004	2.363267e+14	5637240	0	2016-04-29	46	FONTE GRANDE	0	1	0
1013	6.544869e+10	5637358	0	2016-04-29	62	MARUÍPE	0	1	0
1018	3.156533e+14	5596731	1	2016-04-29	53	CENTRO	1	1	0
1035	2.667290e+12	5592387	0	2016-04-29	56	CENTRO	0	1	0
1037	1.753268e+12	5592441	0	2016-04-29	58	CENTRO	0	1	0
1131	4.788217e+14	5544176	1	2016-04-29	38	ROMÃO	1	1	0
1169	1.721570e+13	5570160	1	2016-04-29	31	INHANGUETÁ	0	1	0
1174	6.611178e+13	5602909	1	2016-04-29	85	GRANDE VITÓRIA	0	1	0
1241	5.572283e+13	5617958	0	2016-04-29	40	SÃO PEDRO	0	1	0
...	...	...	...	...	...	...	...	...	...
109748	7.122385e+12	5755458	0	2016-06-03	56	SÃO JOSÉ	0	1	0
109777	7.117382e+10	5768290	0	2016-06-07	80	PRAIA DO SUÁ	0	1	1
109796	5.423193e+10	5755436	0	2016-06-03	57	ESTRELINHA	0	1	0
109802	6.495593e+14	5755449	0	2016-06-03	64	JOANA D´ARC	0	1	0
109886	7.236735e+12	5778542	0	2016-06-08	71	SANTA MARTHA	0	1	0
109892	7.842942e+13	5755231	0	2016-06-03	86	BONFIM	0	1	0
109901	5.334474e+13	5767039	1	2016-06-06	73	CONSOLAÇÃO	0	1	1
109927	1.181477e+13	5773220	1	2016-06-08	76	BONFIM	0	1	1
109948	4.727179e+11	5748877	1	2016-06-02	49	JUCUTUQUARA	0	1	0
109956	3.651334e+13	5768070	0	2016-06-07	94	CARATOÍRA	0	1	1
109961	8.336469e+14	5768062	1	2016-06-07	58	SÃO JOSÉ	0	1	0
110011	5.521582e+14	5770772	1	2016-06-08	61	DO QUADRO	0	1	0
110015	2.522419e+09	5770766	1	2016-06-08	55	CENTRO	0	1	0
110086	5.778776e+12	5734671	1	2016-06-01	72	CENTRO	0	1	0
110116	5.799786e+11	5765617	1	2016-06-06	49	GOIABEIRAS	0	1	0
110129	2.998129e+12	5776686	1	2016-06-08	61	COMDUSA	0	1	1
110142	4.565289e+11	5771110	1	2016-06-07	62	SANTA CLARA	0	1	0
110144	6.213667e+13	5766332	1	2016-06-06	57	PARQUE MOSCOSO	0	1	0
110146	1.961331e+13	5766330	1	2016-06-06	66	DA PENHA	0	1	0
110152	8.946400e+14	5772347	0	2016-06-07	36	VILA RUBIM	0	1	0
110168	5.654627e+12	5756834	0	2016-06-03	50	SANTO ANTÔNIO	0	1	1
110197	9.238845e+12	5767697	1	2016-06-08	65	SÃO CRISTÓVÃO	0	1	0
110201	3.674355e+14	5767692	1	2016-06-07	80	CONSOLAÇÃO	0	1	0
110363	2.123885e+14	5624922	1	2016-06-02	54	RESISTÊNCIA	0	1	0
110383	2.957279e+12	5582577	1	2016-06-01	48	RESISTÊNCIA	0	1	0
110386	2.957279e+12	5582576	1	2016-06-01	48	RESISTÊNCIA	0	1	0

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
110399	9.437123e+13	5692938	1	2016-06-07	17	RESISTÊNCIA	0	1	0
110492	6.456342e+14	5786741	0	2016-06-08	33	MARIA ORTIZ	0	1	0
110496	8.544295e+13	5779046	1	2016-06-08	37	MARIA ORTIZ	0	1	0
110515	6.456342e+14	5778621	0	2016-06-08	33	MARIA ORTIZ	0	1	0

3768 rows × 15 columns

In [48]:

```
# Create variables for diabetes, hypertension positive
hypertension_yes = df['hypertension'] == 1
diabetes_yes = df['diabetes'] == 1
```

In [49]:

```
# Select all cases where hypertension and missed appointment are true
df[missed & hypertension_yes]
```

Out[49]:

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
44	2.741649e+11	5635414	1	2016-04-29	78	SÃO CRISTÓVÃO	0	1	1
212	4.266984e+14	5642059	0	2016-04-29	62	SANTOS DUMONT	0	1	1
270	8.617228e+12	5620528	1	2016-04-29	45	CARATOÍRA	1	1	0
272	5.119616e+12	5594085	1	2016-04-29	51	NOVA PALESTINA	0	1	0
352	5.917359e+12	5494413	1	2016-04-29	62	SÃO CRISTÓVÃO	0	1	0
441	3.935966e+13	5639484	0	2016-04-29	36	SÃO BENEDITO	0	1	0
463	5.228864e+08	5633872	1	2016-04-29	66	SÃO CRISTÓVÃO	0	1	1
537	3.546481e+13	5590085	0	2016-04-29	58	ITARARÉ	0	1	0
619	4.272659e+13	5634178	1	2016-04-29	67	PRAIA DO CANTO	0	1	1
684	8.238132e+13	5629179	1	2016-04-29	73	JOANA D´ARC	0	1	0
729	3.969538e+12	5636130	1	2016-04-29	69	CENTRO	0	1	1
751	6.468148e+14	5629130	0	2016-04-29	68	SÃO PEDRO	0	1	0
780	4.478798e+12	5611940	1	2016-04-29	69	JARDIM DA PENHA	0	1	0
914	1.899995e+13	5525204	0	2016-04-29	64	MARIA ORTIZ	0	1	0
951	7.414865e+12	5317449	1	2016-04-29	77	JESUS DE NAZARETH	0	1	0
968	7.244332e+14	5399572	1	2016-04-29	78	CRUZAMENTO	0	1	1
972	8.941338e+10	5361416	0	2016-04-29	75	BONFIM	0	1	1
973	8.219692e+14	5331088	1	2016-04-29	66	MARIA ORTIZ	0	1	1
974	1.729243e+14	5505270	0	2016-04-29	82	GRANDE VITÓRIA	0	1	0
975	7.584570e+14	5522365	0	2016-04-29	67	MARUÍPE	0	1	0
979	2.539578e+12	5627308	1	2016-04-29	81	SANTO ANDRÉ	0	1	1
1004	2.363267e+14	5637240	0	2016-04-29	46	FONTE GRANDE	0	1	0
1013	6.544869e+10	5637358	0	2016-04-29	62	MARUÍPE	0	1	0

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabete
1018	3.156533e+14	5596731	1	2016-04-29	53	CENTRO	1	1	0
1035	2.667290e+12	5592387	0	2016-04-29	56	CENTRO	0	1	0
1037	1.753268e+12	5592441	0	2016-04-29	58	CENTRO	0	1	0
1131	4.788217e+14	5544176	1	2016-04-29	38	ROMÃO	1	1	0
1169	1.721570e+13	5570160	1	2016-04-29	31	INHANGUETÁ	0	1	0
1174	6.611178e+13	5602909	1	2016-04-29	85	GRANDE VITÓRIA	0	1	0
1241	5.572283e+13	5617958	0	2016-04-29	40	SÃO PEDRO	0	1	0
...	...	...	...	...	...	...	...	...	...
109748	7.122385e+12	5755458	0	2016-06-03	56	SÃO JOSÉ	0	1	0
109777	7.117382e+10	5768290	0	2016-06-07	80	PRAIA DO SUÁ	0	1	1
109796	5.423193e+10	5755436	0	2016-06-03	57	ESTRELINHA	0	1	0
109802	6.495593e+14	5755449	0	2016-06-03	64	JOANA D´ARC	0	1	0
109886	7.236735e+12	5778542	0	2016-06-08	71	SANTA MARTHA	0	1	0
109892	7.842942e+13	5755231	0	2016-06-03	86	BONFIM	0	1	0
109901	5.334474e+13	5767039	1	2016-06-06	73	CONSOLAÇÃO	0	1	1
109927	1.181477e+13	5773220	1	2016-06-08	76	BONFIM	0	1	1
109948	4.727179e+11	5748877	1	2016-06-02	49	JUCUTUQUARA	0	1	0
109956	3.651334e+13	5768070	0	2016-06-07	94	CARATOÍRA	0	1	1
109961	8.336469e+14	5768062	1	2016-06-07	58	SÃO JOSÉ	0	1	0
110011	5.521582e+14	5770772	1	2016-06-08	61	DO QUADRO	0	1	0
110015	2.522419e+09	5770766	1	2016-06-08	55	CENTRO	0	1	0
110086	5.778776e+12	5734671	1	2016-06-01	72	CENTRO	0	1	0
110116	5.799786e+11	5765617	1	2016-06-06	49	GOIABEIRAS	0	1	0
110129	2.998129e+12	5776686	1	2016-06-08	61	COMDUSA	0	1	1
110142	4.565289e+11	5771110	1	2016-06-07	62	SANTA CLARA	0	1	0
110144	6.213667e+13	5766332	1	2016-06-06	57	PARQUE MOSCOSO	0	1	0
110146	1.961331e+13	5766330	1	2016-06-06	66	DA PENHA	0	1	0
110152	8.946400e+14	5772347	0	2016-06-07	36	VILA RUBIM	0	1	0
110168	5.654627e+12	5756834	0	2016-06-03	50	SANTO ANTÔNIO	0	1	1
110197	9.238845e+12	5767697	1	2016-06-08	65	SÃO CRISTÓVÃO	0	1	0
110201	3.674355e+14	5767692	1	2016-06-07	80	CONSOLAÇÃO	0	1	0
110363	2.123885e+14	5624922	1	2016-06-02	54	RESISTÊNCIA	0	1	0
110383	2.957279e+12	5582577	1	2016-06-01	48	RESISTÊNCIA	0	1	0
110386	2.957279e+12	5582576	1	2016-06-01	48	RESISTÊNCIA	0	1	0
110399	9.437123e+13	5692938	1	2016-06-07	17	RESISTÊNCIA	0	1	0
110492	6.456342e+14	5786741	0	2016-06-08	33	MARIA ORTIZ	0	1	0
110496	8.544295e+13	5779046	1	2016-06-08	37	MARIA ORTIZ	0	1	0
110515	6.456342e+14	5778621	0	2016-06-08	33	MARIA ORTIZ	0	1	0

3768 rows × 15 columns

```
In [50]:
```

```
#Select all cases where diabetes and missed appointment are true
df[missed & diabetes_yes]
```

Out[50]:

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
44	2.741649e+11	5635414	1	2016-04-29	78	SÃO CRISTÓVÃO	0	1	1
126	9.447582e+14	5633576	1	2016-04-29	67	PRAIA DO SUÁ	0	0	1
212	4.266984e+14	5642059	0	2016-04-29	62	SANTOS DUMONT	0	1	1
463	5.228864e+08	5633872	1	2016-04-29	66	SÃO CRISTÓVÃO	0	1	1
619	4.272659e+13	5634178	1	2016-04-29	67	PRAIA DO CANTO	0	1	1
729	3.969538e+12	5636130	1	2016-04-29	69	CENTRO	0	1	1
968	7.244332e+14	5399572	1	2016-04-29	78	CRUZAMENTO	0	1	1
972	8.941338e+10	5361416	0	2016-04-29	75	BONFIM	0	1	1
973	8.219692e+14	5331088	1	2016-04-29	66	MARIA ORTIZ	0	1	1
979	2.539578e+12	5627308	1	2016-04-29	81	SANTO ANDRÉ	0	1	1
1276	8.655848e+14	5507702	1	2016-04-29	15	JOANA D´ARC	0	0	1
1430	3.656627e+14	5612155	1	2016-04-29	67	VILA RUBIM	0	1	1
1471	1.957563e+12	5602435	0	2016-04-29	74	CARATOÍRA	0	0	1
1608	3.856226e+12	5589516	0	2016-04-29	59	TABUAZEIRO	0	1	1
1625	8.472339e+11	5618144	1	2016-04-29	69	GURIGICA	0	1	1
1626	7.381976e+12	5618145	1	2016-04-29	74	GURIGICA	0	1	1
1627	3.444299e+10	5618132	0	2016-04-29	40	GURIGICA	0	1	1
1628	8.867617e+12	5618128	1	2016-04-29	50	GURIGICA	0	1	1
1630	8.423915e+12	5618140	1	2016-04-29	60	GURIGICA	0	1	1
1640	8.513189e+12	5618148	0	2016-04-29	63	GURIGICA	0	1	1
1642	2.495740e+14	5618142	0	2016-04-29	48	CONSOLAÇÃO	0	1	1
1643	8.645720e+14	5618137	1	2016-04-29	63	GURIGICA	0	1	1
1644	7.385364e+12	5618149	1	2016-04-29	59	GURIGICA	1	1	1
1646	3.593727e+14	5618139	1	2016-04-29	68	GURIGICA	0	1	1
1778	3.649915e+12	5536651	1	2016-04-29	60	DO MOSCOSO	1	1	1
1999	4.325445e+13	5351222	1	2016-04-29	55	RESISTÊNCIA	0	1	1
2255	5.944667e+13	5451793	1	2016-04-29	49	SÃO BENEDITO	1	1	1
2380	8.875659e+13	5622850	1	2016-04-29	80	ANDORINHAS	0	1	1
2395	1.766491e+14	5630786	1	2016-04-29	74	SÃO JOSÉ	0	1	1
2399	5.746559e+14	5626765	1	2016-04-29	48	SÃO JOSÉ	0	1	1
...	...	...	...	...	...	...	...	...	...
107719	3.911518e+14	5768338	1	2016-06-07	58	SANTO ANTÔNIO	0	1	1
107733	1.171288e+14	5743290	1	2016-06-01	57	REDENÇÃO	0	1	1
107740	5.837441e+13	5743260	0	2016-06-01	51	SANTA CLARA	0	1	1
107864	8.996672e+14	5562304	1	2016-06-01	77	DE LOURDES	0	1	1
						NOVA			

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
108005	9.164578e+10	5762605	1	2016-06-02	60	NOVA PALESTINA	1	1	1
108165	8.317640e+14	5520974	1	2016-06-07	51	CRUZAMENTO	0	1	1
108166	9.326460e+14	5520979	1	2016-06-07	57	CRUZAMENTO	0	1	1
108169	4.974497e+13	5520978	0	2016-06-07	62	CRUZAMENTO	0	1	1
108344	7.937375e+13	5762094	1	2016-06-08	44	ROMÃO	0	1	1
108406	4.795643e+12	5716046	1	2016-06-01	85	CARATOÍRA	0	1	1
108411	1.957563e+12	5716037	0	2016-06-01	74	CARATOÍRA	0	0	1
108513	1.957344e+12	5745266	1	2016-06-01	74	DE LOURDES	0	1	1
108854	9.677914e+12	5745629	1	2016-06-01	52	GURIGICA	0	1	1
108922	6.166587e+14	5761590	0	2016-06-06	68	FONTE GRANDE	0	1	1
109397	9.922630e+13	5748354	0	2016-06-01	62	BENTO FERREIRA	0	1	1
109423	2.578499e+13	5767076	1	2016-06-06	63	SÃO JOSÉ	0	1	1
109479	1.842445e+14	5752466	0	2016-06-02	67	CONQUISTA	0	1	1
109494	8.164175e+13	5778716	1	2016-06-08	40	MONTE BELO	0	0	1
109503	3.619493e+12	5745207	1	2016-06-01	49	RESISTÊNCIA	0	0	1
109506	3.836521e+14	5778690	1	2016-06-08	67	RESISTÊNCIA	0	1	1
109585	7.582520e+12	5772435	0	2016-06-07	52	COMDUSA	0	0	1
109593	2.436681e+13	5766725	0	2016-06-06	53	RESISTÊNCIA	0	1	1
109607	5.546174e+12	5753653	1	2016-06-01	54	BENTO FERREIRA	0	1	1
109777	7.117382e+10	5768290	0	2016-06-07	80	PRAIA DO SUÁ	0	1	1
109901	5.334474e+13	5767039	1	2016-06-06	73	CONSOLAÇÃO	0	1	1
109927	1.181477e+13	5773220	1	2016-06-08	76	BONFIM	0	1	1
109956	3.651334e+13	5768070	0	2016-06-07	94	CARATOÍRA	0	1	1
110062	3.915317e+12	5741991	1	2016-06-02	42	TABUAZEIRO	1	0	1
110129	2.998129e+12	5776686	1	2016-06-08	61	COMDUSA	0	1	1
110168	5.654627e+12	5756834	0	2016-06-03	50	SANTO ANTÔNIO	0	1	1

1430 rows × 15 columns

In [51]:

```python
#select all rows where the patient was absent and has diabetes
df[(df['diabetes'] == 1) & (df['absenteeism'] == 1)]
```

Out[51]:

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
44	2.741649e+11	5635414	1	2016-04-29	78	SÃO CRISTÓVÃO	0	1	1
126	9.447582e+14	5633576	1	2016-04-29	67	PRAIA DO SUÁ	0	0	1
212	4.266984e+14	5642059	0	2016-04-29	62	SANTOS DUMONT	0	1	1
463	5.228864e+08	5633872	1	2016-04-29	66	SÃO CRISTÓVÃO	0	1	1
619	4.272659e+13	5634178	1	2016-04-29	67	PRAIA DO CANTO	0	1	1

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
729	3.969538e+12	5636130	1	2016-04-29	69	CENTRO	0	1	1
968	7.244332e+14	5399572	1	2016-04-29	78	CRUZAMENTO	0	1	1
972	8.941338e+10	5361416	0	2016-04-29	75	BONFIM	0	1	1
973	8.219692e+14	5331088	1	2016-04-29	66	MARIA ORTIZ	0	1	1
979	2.539578e+12	5627308	1	2016-04-29	81	SANTO ANDRÉ	0	1	1
1276	8.655848e+14	5507702	1	2016-04-29	15	JOANA D´ARC	0	0	1
1430	3.656627e+14	5612155	1	2016-04-29	67	VILA RUBIM	0	1	1
1471	1.957563e+12	5602435	0	2016-04-29	74	CARATOÍRA	0	0	1
1608	3.856226e+12	5589516	0	2016-04-29	59	TABUAZEIRO	0	1	1
1625	8.472339e+11	5618144	1	2016-04-29	69	GURIGICA	0	1	1
1626	7.381976e+12	5618145	1	2016-04-29	74	GURIGICA	0	1	1
1627	3.444299e+10	5618132	0	2016-04-29	40	GURIGICA	0	1	1
1628	8.867617e+12	5618128	1	2016-04-29	50	GURIGICA	0	1	1
1630	8.423915e+12	5618140	1	2016-04-29	60	GURIGICA	0	1	1
1640	8.513189e+12	5618148	0	2016-04-29	63	GURIGICA	0	1	1
1642	2.495740e+14	5618142	0	2016-04-29	48	CONSOLAÇÃO	0	1	1
1643	8.645720e+14	5618137	1	2016-04-29	63	GURIGICA	0	1	1
1644	7.385364e+12	5618149	1	2016-04-29	59	GURIGICA	1	1	1
1646	3.593727e+14	5618139	1	2016-04-29	68	GURIGICA	0	1	1
1778	3.649915e+12	5536651	1	2016-04-29	60	DO MOSCOSO	1	1	1
1999	4.325445e+13	5351222	1	2016-04-29	55	RESISTÊNCIA	0	1	1
2255	5.944667e+13	5451793	1	2016-04-29	49	SÃO BENEDITO	1	1	1
2380	8.875659e+13	5622850	1	2016-04-29	80	ANDORINHAS	0	1	1
2395	1.766491e+14	5630786	1	2016-04-29	74	SÃO JOSÉ	0	1	1
2399	5.746559e+14	5626765	1	2016-04-29	48	SÃO JOSÉ	0	1	1
...	...	...	...	...	...	...	...	...	...
107719	3.911518e+14	5768338	1	2016-06-07	58	SANTO ANTÔNIO	0	1	1
107733	1.171288e+14	5743290	1	2016-06-01	57	REDENÇÃO	0	1	1
107740	5.837441e+13	5743260	0	2016-06-01	51	SANTA CLARA	0	1	1
107864	8.996672e+14	5562304	1	2016-06-01	77	DE LOURDES	0	1	1
108005	9.164578e+10	5762605	1	2016-06-02	60	NOVA PALESTINA	1	1	1
108165	8.317640e+14	5520974	1	2016-06-07	51	CRUZAMENTO	0	1	1
108166	9.326460e+14	5520979	1	2016-06-07	57	CRUZAMENTO	0	1	1
108169	4.974497e+13	5520978	0	2016-06-07	62	CRUZAMENTO	0	1	1
108344	7.937375e+13	5762094	1	2016-06-08	44	ROMÃO	0	1	1
108406	4.795643e+12	5716046	1	2016-06-01	85	CARATOÍRA	0	1	1
108411	1.957563e+12	5716037	0	2016-06-01	74	CARATOÍRA	0	0	1
108513	1.957344e+12	5745266	1	2016-06-01	74	DE LOURDES	0	1	1
108854	9.677914e+12	5745629	1	2016-06-01	52	GURIGICA	0	1	1
108922	6.166587e+14	5761590	0	2016-06-06	68	FONTE GRANDE	0	1	1
109397	9.922630e+13	5748354	0	2016-06-01	62	BENTO FERREIRA	0	1	1

	patient_id	appointment_id	gender	appointment_day	age	neighborhood	scholarship	hypertension	diabetes
109423	2.578499e+13	5767076	1	2016-06-06	63	SÃO JOSÉ	0	1	1
109479	1.842445e+14	5752466	0	2016-06-02	67	CONQUISTA	0	1	1
109494	8.164175e+13	5778716	1	2016-06-08	40	MONTE BELO	0	0	1
109503	3.619493e+12	5745207	1	2016-06-01	49	RESISTÊNCIA	0	0	1
109506	3.836521e+14	5778690	1	2016-06-08	67	RESISTÊNCIA	0	1	1
109585	7.582520e+12	5772435	0	2016-06-07	52	COMDUSA	0	0	1
109593	2.436681e+13	5766725	0	2016-06-06	53	RESISTÊNCIA	0	1	1
109607	5.546174e+12	5753653	1	2016-06-01	54	BENTO FERREIRA	0	1	1
109777	7.117382e+10	5768290	0	2016-06-07	80	PRAIA DO SUÁ	0	1	1
109901	5.334474e+13	5767039	1	2016-06-06	73	CONSOLAÇÃO	0	1	1
109927	1.181477e+13	5773220	1	2016-06-08	76	BONFIM	0	1	1
109956	3.651334e+13	5768070	0	2016-06-07	94	CARATOÍRA	0	1	1
110062	3.915317e+12	5741991	1	2016-06-02	42	TABUAZEIRO	1	0	1
110129	2.998129e+12	5776686	1	2016-06-08	61	COMDUSA	0	1	1
110168	5.654627e+12	5756834	0	2016-06-03	50	SANTO ANTÔNIO	0	1	1

1430 rows × 15 columns

In [52]:

```
#Groupby hyptension and absenteeism and see a descriptive summary
df.groupby(['hypertension', 'absenteeism']).describe()
```

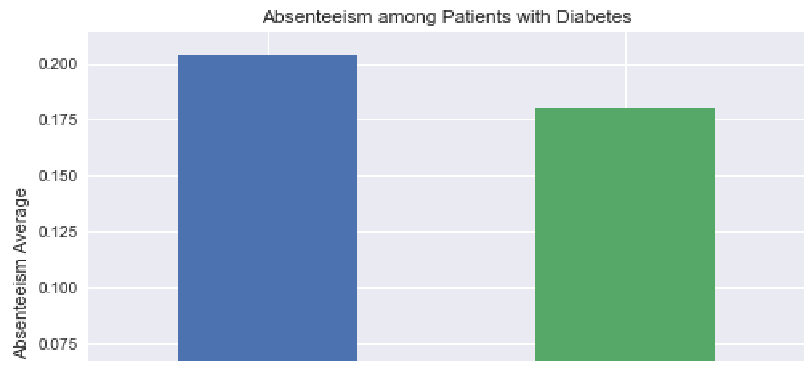
Out[52]:

		age								alcoholism		...	sms_received	
		count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**count**	**mean**	**...**	**75%**	**max**
hypertension	**absenteeism**													
0	**0**	70163.0	31.864772	21.534238	0.0	14.0	31.0	49.0	95.0	70163.0	0.022804	...	1.0	1.0
	1	18544.0	29.137349	19.476148	0.0	14.0	27.0	43.0	95.0	18544.0	0.023350	...	1.0	1.0
1	**0**	18005.0	60.751291	13.762684	7.0	52.0	61.0	70.0	95.0	18005.0	0.060150	...	1.0	1.0
	1	3768.0	59.681529	14.369464	4.0	51.0	59.0	69.0	95.0	3768.0	0.064756	...	1.0	1.0

4 rows × 80 columns

In [53]:

```
df.groupby('diabetes').absenteeism.mean().plot(kind='bar')
plt.title('Absenteeism among Patients with Diabetes')
plt.xlabel("Absenteeism Review: arrived vs. missed")
plt.ylabel("Absenteeism Average" )
plt.show();
```

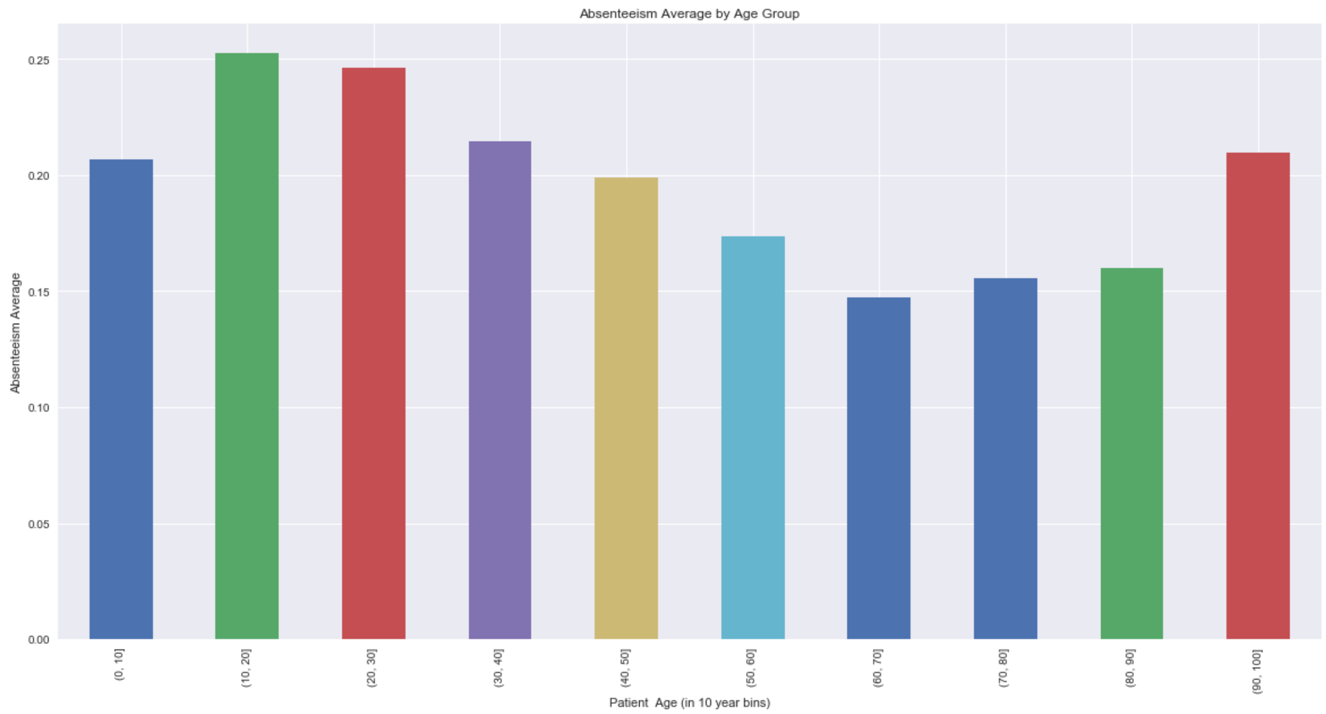


3. What is the relationship between absenteeism and age?

> **Age & Absenteeism:** Appointment_ids held by Patients aged 10-20 are on average most likely to end up missed. By count, appointment_ids held by patients aged 50-60 are most likely to end up missed.

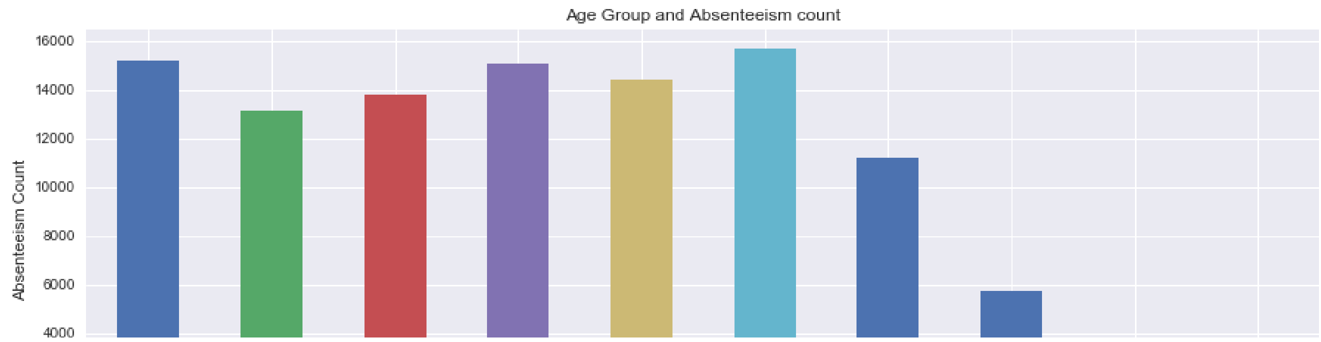
In [54]:

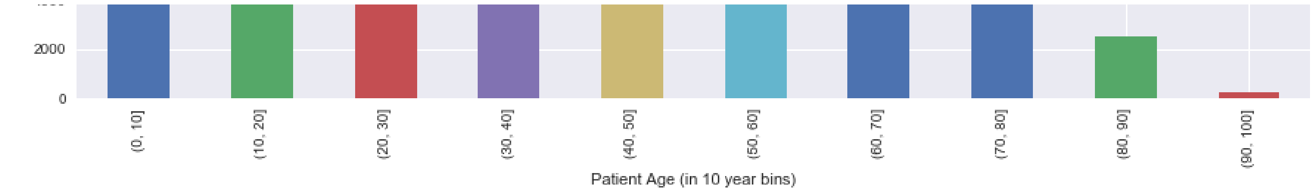
```
df.groupby('age_bins').absenteeism.mean().plot(kind='bar',figsize=(20, 10))
plt.title('Absenteeism Average by Age Group')
plt.xlabel("Patient  Age (in 10 year bins)")
plt.ylabel("Absenteeism Average" )
plt.show()
```



In [55]:

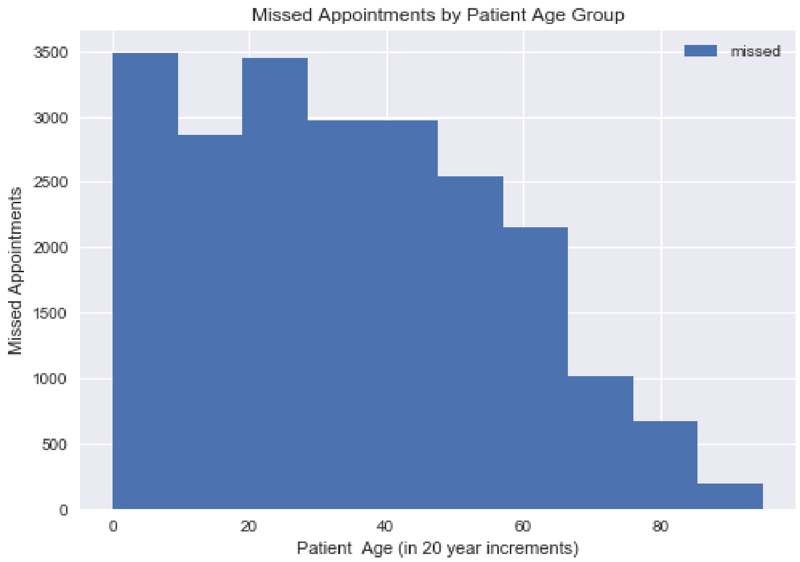
```
df_age = df[['absenteeism', 'age_bins']].groupby('age_bins').count()
plot_title = 'Age Group and Absenteeism count'
ax = df_age['absenteeism'].plot(kind='bar', figsize=(15, 5), title=plot_title)
ax.set_ylabel('Absenteeism Count')
ax.set_xlabel('Patient Age (in 10 year bins)');
```





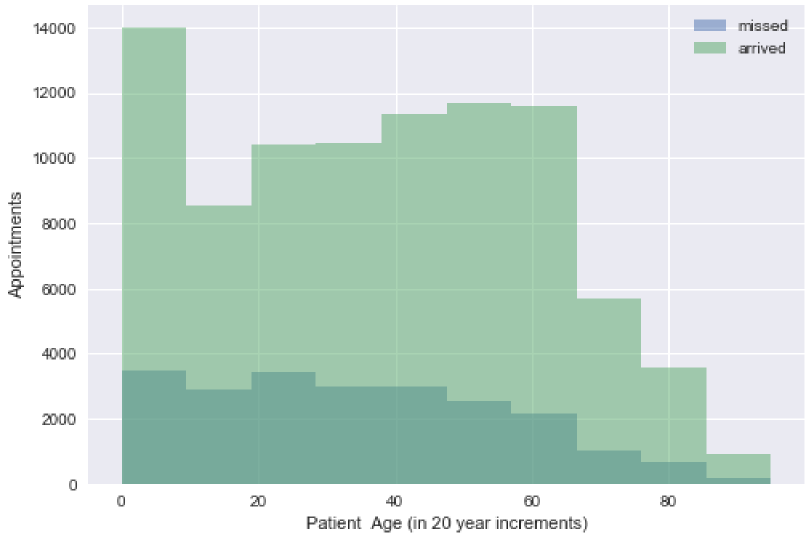
In [56]:

```
df.age[missed].hist(label='missed');
plt.title('Missed Appointments by Patient Age Group')
plt.xlabel("Patient  Age (in 20 year increments)")
plt.ylabel("Missed Appointments" )
plt.legend();
```



In [57]:

```
df.age[missed].hist(alpha=0.5, label='missed')
df.age[arrived].hist(alpha=0.5, label='arrived')
plt.xlabel("Patient  Age (in 20 year increments)")
plt.ylabel("Appointments" )
plt.legend();
```



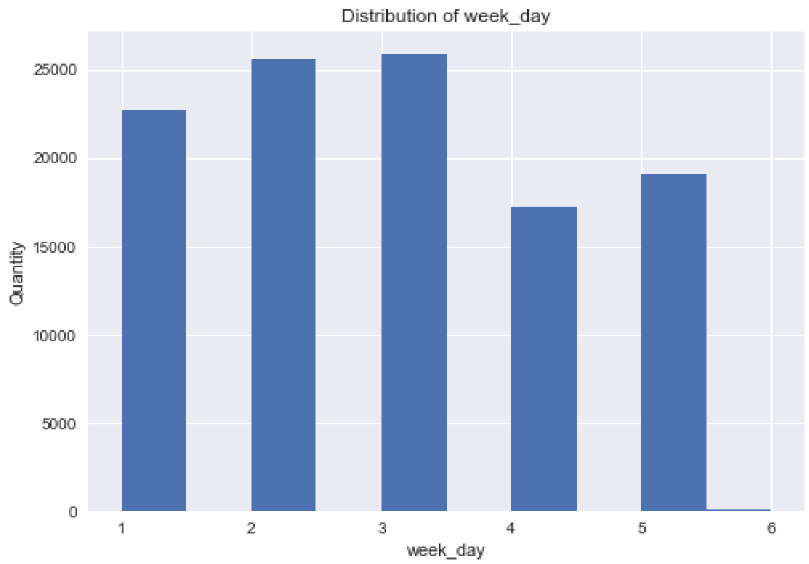
4. What is the relationship between absenteeism and appointment day?

> **Day of the Week & Absenteeism:** Appointments are more likely to be missed on Tuesdays.

Day of the Week & Absenteeism: Appointments are more likely to be missed on Tuesdays.

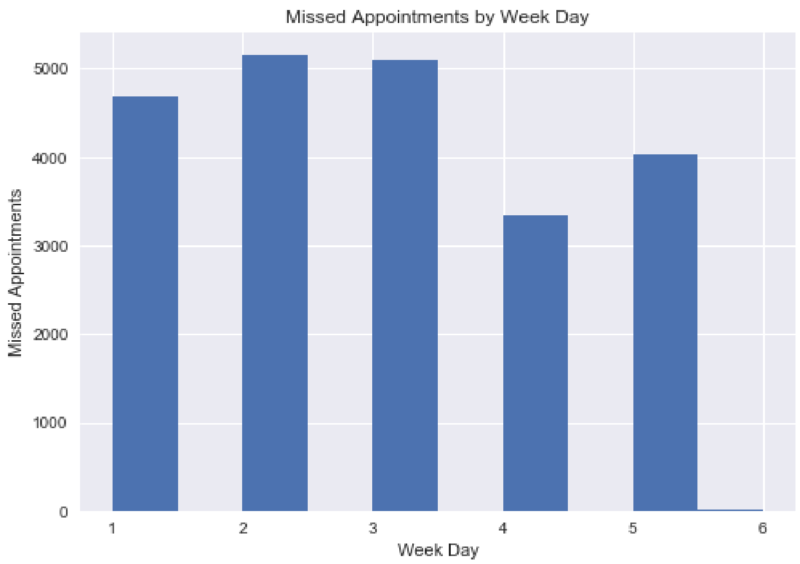
In [58]:

```
# Distribution of 'week_day'
plt.figure();
age_hist = df['week_day'].plot.hist(bins=10)
age_hist.set_xlabel("week_day")
age_hist.set_ylabel("Quantity")
age_hist.set_title('Distribution of week_day');
```



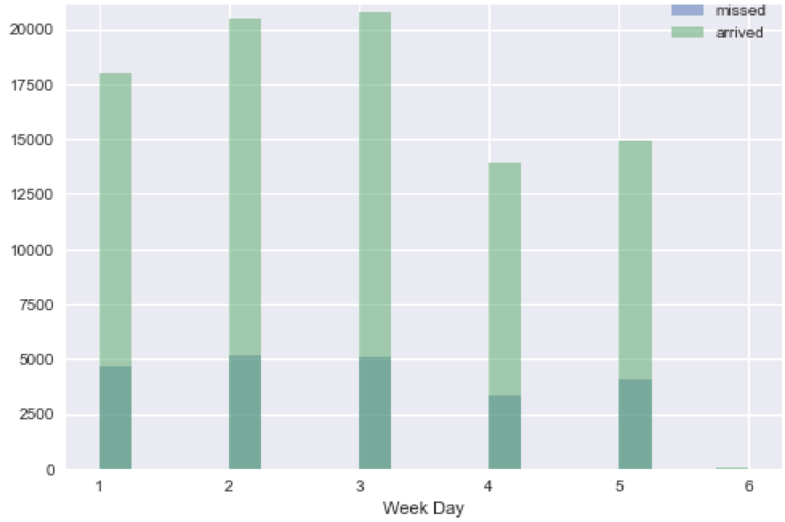
In [59]:

```
df.week_day[missed].hist(label='missed')
plt.title('Missed Appointments by Week Day')
plt.xlabel("Week Day")
plt.ylabel("Missed Appointments" )
plt.show();
```



In [60]:

```
df.week_day[missed].hist(alpha=0.5, bins = 20, label='missed')
df.week_day[arrived].hist(alpha=0.5, bins = 20, label='arrived')
plt.title('Appointments by Week Day')
plt.xlabel("Week Day")
plt.legend();
```



Conclusions

- A patient's age appears to have an impact on likelihood to miss appointments.
- SMS texts received appear to have a negative impact on keeping appointments.
- Certain days have higher missed appointment rates

Limitations of dataset

More information is needed on the conditions for which SMS texts were sent to patients.

In [61]:

```
from subprocess import call
call(['python', '-m', 'nbconvert', 'Missed_Appointments.ipynb'])
```

Out[61]:

```
255
```

In [ ]:

```
import pdfkit
pdfkit.from_file('Missed_Appointments.html', 'missed_appointments_analysis.pdf')
```