Wrangling and Visualizing Alzheimer Data with Plotly

Dataset from https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset/data

In [22]:
```
# Import Libraries
import pandas as pd
import plotly.express as px
import plotly
plotly.offline.init_notebook_mode(connected=True)
import numpy as np
```

In [23]:
```
# Read in File
df = pd.read_csv("/content/alzheimers_disease_data.csv")
df.head()
```

Out[23]:

	PatientID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	Alcoho
0	4751	73	0	0	2	22.927749	0	
1	4752	89	0	0	0	26.827681	0	
2	4753	73	0	3	1	17.795882	0	
3	4754	74	1	0	1	33.800817	1	
4	4755	89	0	0	0	20.716974	0	

5 rows × 35 columns

In [24]:
```
# Explore Data
df.info()
```

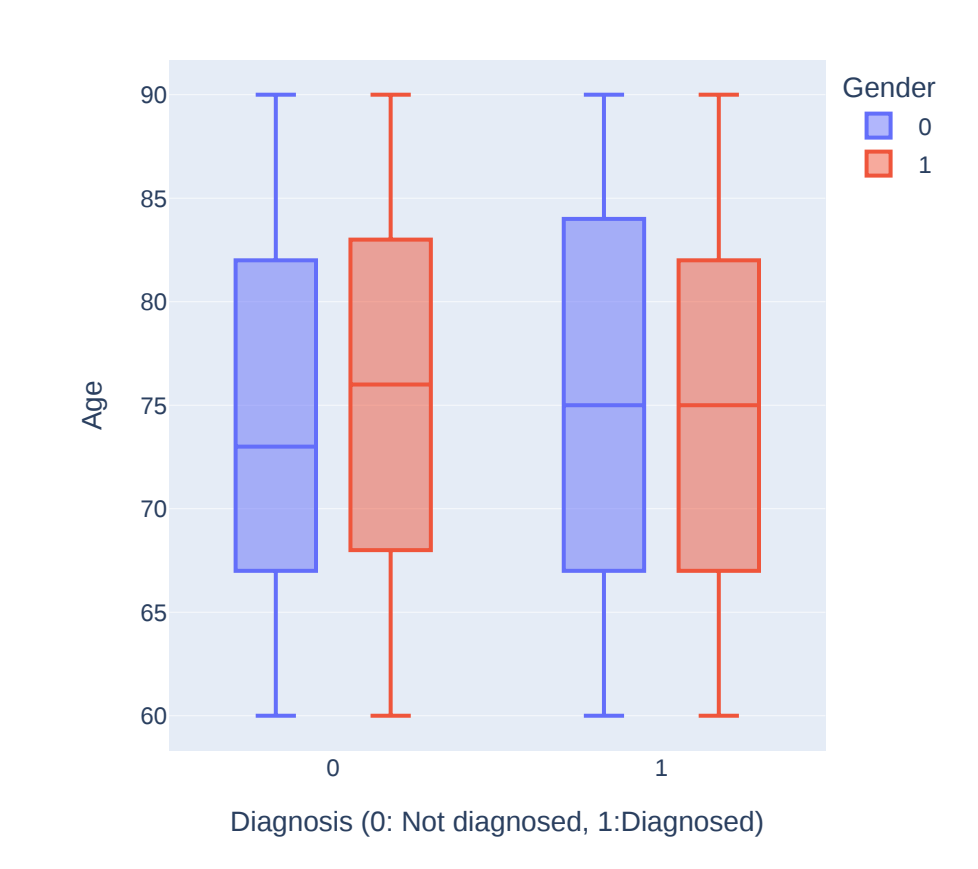
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2149 entries, 0 to 2148
Data columns (total 35 columns):
 #   Column                     Non-Null Count  Dtype  
---  ------                     --------------  -----  
 0   PatientID                  2149 non-null   int64  
 1   Age                        2149 non-null   int64  
 2   Gender                     2149 non-null   int64  
 3   Ethnicity                  2149 non-null   int64  
 4   EducationLevel             2149 non-null   int64  
 5   BMI                        2149 non-null   float64
 6   Smoking                    2149 non-null   int64  
 7   AlcoholConsumption         2149 non-null   float64
 8   PhysicalActivity           2149 non-null   float64
 9   DietQuality                2149 non-null   float64
 10  SleepQuality               2149 non-null   float64
 11  FamilyHistoryAlzheimers    2149 non-null   int64  
 12  CardiovascularDisease      2149 non-null   int64  
 13  Diabetes                   2149 non-null   int64  
 14  Depression                 2149 non-null   int64  
 15  HeadInjury                 2149 non-null   int64  
 16  Hypertension               2149 non-null   int64  
 17  SystolicBP                 2149 non-null   int64  
 18  DiastolicBP                2149 non-null   int64  
 19  CholesterolTotal           2149 non-null   float64
 20  CholesterolLDL             2149 non-null   float64
 21  CholesterolHDL             2149 non-null   float64
 22  CholesterolTriglycerides   2149 non-null   float64
 23  MMSE                       2149 non-null   float64
 24  FunctionalAssessment       2149 non-null   float64
 25  MemoryComplaints           2149 non-null   int64  
 26  BehavioralProblems         2149 non-null   int64  
 27  ADL                        2149 non-null   float64
 28  Confusion                  2149 non-null   int64  
 29  Disorientation             2149 non-null   int64  
 30  PersonalityChanges         2149 non-null   int64  
 31  DifficultyCompletingTasks  2149 non-null   int64  
 32  Forgetfulness              2149 non-null   int64  
 33  Diagnosis                  2149 non-null   int64  
 34  DoctorInCharge             2149 non-null   object 
dtypes: float64(12), int64(22), object(1)
memory usage: 587.7+ KB
```

Demographic Details in dataset

- **Age**: The age of the patients ranges from 60 to 90 years.
- **Gender**: Gender of the patients, where 0 represents Male and 1 represents Female.
- **Ethnicity**: The ethnicity of the patients, coded as follows: 0: Caucasian 1: African American 2: Asian 3: Other
- **EducationLevel**: The education level of the patients, coded as follows: 0: None 1: High School 2: Bachelor's 3: Higher

In [26]:
```
# Visualization with Age
fig1 = px.box(
    df,
    x = "Diagnosis",
    y = "Age",
    color = "Gender",
    title="Correlation between Age and Alzheimer Diagnosis, Facet
    labels={"Age": "Age",
            "Diagnosis": "Diagnosis (0: Not diagnosed, 1:Diagnose
)
fig1.show()
```

Correlation between Age and Alzheimer Diagnosis, Facet by G



In [27]:
```
# Wrangling with Gender

count_male = (
    df
    .query("Gender == 0")
    .shape[0]
)
count_female = (
    df
    .query("Gender == 1")
    .shape[0]
)
```

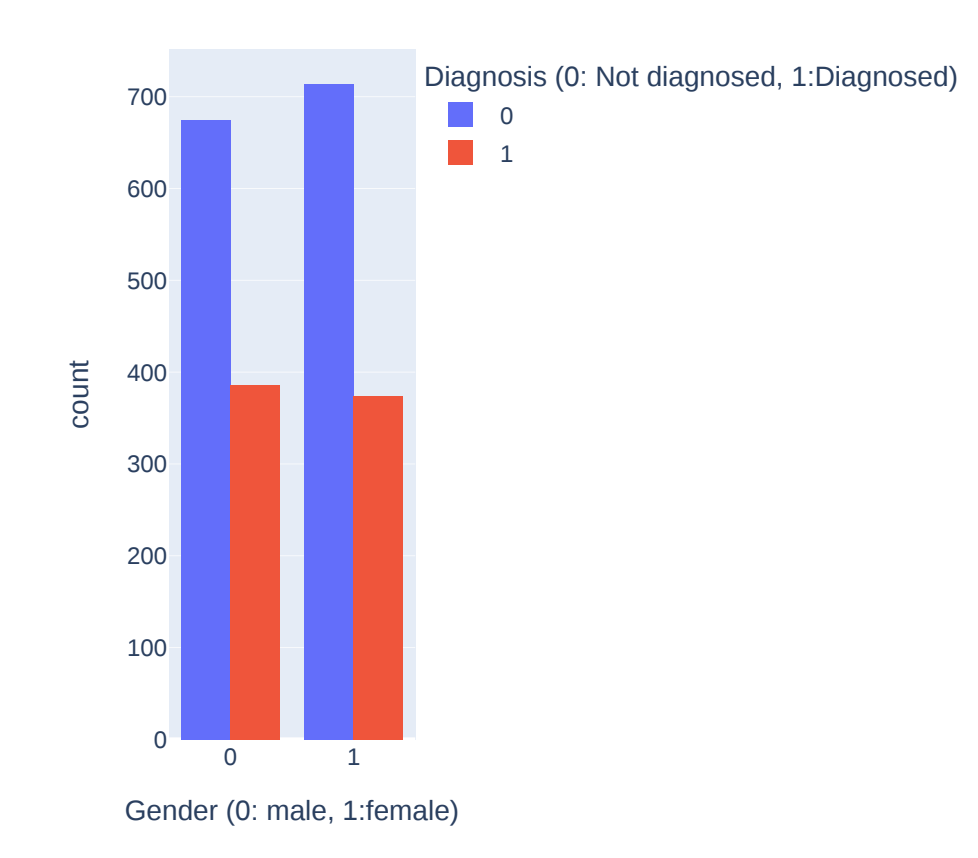
In [28]:
```
Alz_by_gender = (
    df
    .groupby("Gender")
    .agg(sum_diagnosis = ('Diagnosis','sum'))
)
Alz_by_gender["gender count"] = [count_male, count_female]
Alz_by_gender
```

Out[28]:

Gender	sum_diagnosis	gender count
0	386	1061
1	374	1088

In [29]:
```
# visualization
fig = px.histogram(df,
                   x = "Gender",
                   color = "Diagnosis",
                   barmode = "group",
                   title = "Diagnosis grouped by gender",
                   labels = {"Gender": "Gender (0: male, 1:female
                             "Diagnosis": "Diagnosis (0: Not diag
                   )
fig.show()
```

Diagnosis grouped by gender



Lifestyle Factors

- **BMI**: Body Mass Index of the patients, ranging from 15 to 40.
- **Smoking**: Smoking status, where 0 indicates No and 1 indicates Yes.
- **AlcoholConsumption**: Weekly alcohol consumption in units, ranging from 0 to 20.
- **PhysicalActivity**: Weekly physical activity in hours, ranging from 0 to 10.
- **DietQuality**: Diet quality score, ranging from 0 to 10.
- **SleepQuality**: Sleep quality score, ranging from 4 to 10.

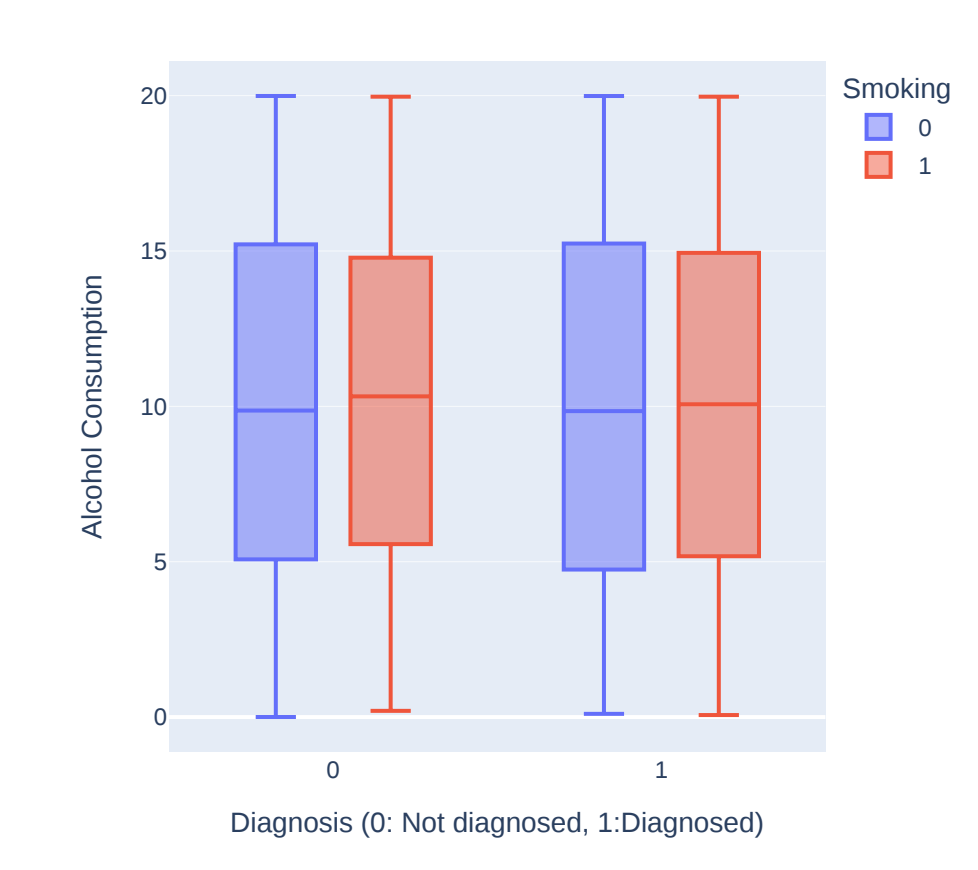
In [30]:
```
Alz_by_lf = (
    df
    .groupby(["Smoking"], as_index = False)
    .agg(sum_diagnosis = ('Diagnosis','sum'),
         mean_bmi = ('BMI','mean'),
         mean_alcohol = ('AlcoholConsumption','mean'),
         mean_physical = ('PhysicalActivity','mean'),
         mean_diet = ('DietQuality','mean'),
         mean_sleep = ('SleepQuality','mean'))
)
Alz_by_lf
```

Out[30]:

	Smoking	sum_diagnosis	mean_bmi	mean_alcohol	mean_physical	mean_diet
0	0	543	27.561790	10.008784	4.900629	4.996931
1	1	217	27.887284	10.115048	4.968471	4.983784

In [31]:
```
# Visualization correlation between Alcohol Consumption, smoking,
fig2 = px.box(
    df,
    x = "Diagnosis",
    y = "AlcoholConsumption",
    color = "Smoking",
    title="Correlation between Dianosis and Alcohol Consumption",
    labels={"AlcoholConsumption": "Alcohol Consumption",
            "Diagnosis": "Diagnosis (0: Not diagnosed, 1:Diagnose
)
fig2.show()
```

Correlation between Dianosis and Alcohol Consumption



Medical History

- **FamilyHistoryAlzheimers**: Family history of Alzheimer's Disease, where 0 indicates No and 1 indicates Yes.
- **CardiovascularDisease**: Presence of cardiovascular disease, where 0 indicates No and 1 indicates Yes.
- **Diabetes**: Presence of diabetes, where 0 indicates No and 1 indicates Yes.
- **Depression**: Presence of depression, where 0 indicates No and 1 indicates Yes.
- **HeadInjury**: History of head injury, where 0 indicates No and 1 indicates Yes.
- **Hypertension**: Presence of hypertension, where 0 indicates No and 1 indicates Yes.

In [32]:
```
Alz_by_FamilyHistory = df.groupby(["FamilyHistoryAlzheimers"], as
Alz_by_FamilyHistory
```

Out[32]:

	FamilyHistoryAlzheimers	percentage_diagnosis
0	0	0.362788
1	1	0.326568

a negative correlation between Family History and Alzheimers is suggested

In [33]:
```
Alz_by_CVD = df.groupby(["CardiovascularDisease"], as_index = Fal
Alz_by_CVD
```

Out[33]:

	CardiovascularDisease	percentage_diagnosis
0	0	0.347471
1	1	0.390323

A positive correlation between Cardiovascular disease and Alzheimer disease is suggested.

In [34]:
```
Alz_by_Diabetes = df.groupby(["Diabetes"], as_index = False).agg
Alz_by_Diabetes
```

Out[34]:

	Diabetes	percentage_diagnosis
0	0	0.360000
1	1	0.317901

A negative correlation between diabetes and Alzheimer disease is suggested

In [35]:
```
Alz_by_Depression = df.groupby(["Depression"], as_index = False)
Alz_by_Depression
```

Out[35]:

	Depression	percentage_diagnosis
0	0	0.355064
1	1	0.348028

A slight negative correlation between depression and Alzheimer's Disease is suggested. Further analysis needed to determine its significance.

In [36]:
```
Alz_by_HeadInjury = df.groupby(["HeadInjury"], as_index = False)
Alz_by_HeadInjury
```

Out[36]:

	HeadInjury	percentage_diagnosis
0	0	0.356923
1	1	0.321608

A negative correlation between head injury and Alzheimer's Disease is suggested.

In [37]:
```
Alz_by_Hypertension = df.groupby(["Hypertension"], as_index = Fal
Alz_by_Hypertension
```

Out[37]:

	Hypertension	percentage_diagnosis
0	0	0.346638
1	1	0.393750

A positive correlation between Hypertention and Alzheimer's Disease is suggested.

In [38]:
```
# Show interaction between factors.
Alz_by_med = (df
              .groupby(["FamilyHistoryAlzheimers","Cardiovascula
              .agg(percentage_diagnosis = ('Diagnosis','mean'))
              )
Alz_by_med.sort_values(by = "percentage_diagnosis", ascending = F
```

Out[38]:

	FamilyHistoryAlzheimers	CardiovascularDisease	Diabetes	Depression	HeadIn
49	1	1	1	1	
20	0	1	0	1	
26	0	1	1	1	
23	0	1	1	0	
48	1	1	1	0	
25	0	1	1	1	
45	1	1	0	1	
13	0	0	1	1	
3	0	0	0	0	
6	0	0	0	1	
44	1	1	0	0	
24	0	1	1	0	
16	0	1	0	0	
35	1	0	1	0	
12	0	0	1	1	
10	0	0	1	0	
1	0	0	0	0	
41	1	1	0	0	
36	1	0	1	0	
15	0	1	0	0	
22	0	1	1	0	
0	0	0	0	0	
5	0	0	0	1	
2	0	0	0	0	
28	1	0	0	0	
27	1	0	0	0	
19	0	1	0	1	
46	1	1	0	1	
38	1	0	1	1	
4	0	0	0	1	
9	0	0	1	0	
32	1	0	0	1	
31	1	0	0	1	
34	1	0	1	0	
29	1	0	0	0	
8	0	0	1	0	
14	0	0	1	1	
18	0	1	0	0	
21	0	1	0	1	
17	0	1	0	0	
37	1	0	1	0	
39	1	0	1	1	
40	1	0	1	1	
11	0	0	1	0	
42	1	1	0	0	
43	1	1	0	0	
33	1	0	0	1	
30	1	0	0	0	
7	0	0	0	1	
47	1	1	0	1	

People with combination of multiple diseases are more likly to have Alzheimer's Disease.

In [ ]: