```
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import plotly.express as px
import pandas as pd
import sqlite3


##Se inicia la carga de datos##
df_2017 = pd.read_csv('Titulados_Educacion_Superior_2017.csv', delimiter=';')
df_2018 = pd.read_csv('Titulados_Educacion_Superior_2018.csv', delimiter=';')
df_2019 = pd.read_csv('Titulados_Educacion_Superior_2019.csv', delimiter=';')
df_2020 = pd.read_csv('Titulados_Educacion_Superior_2020.csv', delimiter=';')
df_bd = pd.read_excel('base_indices_2017_2021.xlsx')
```

```
/usr/local/lib/python3.7/dist-packages/IPython/core/interactiveshell.py:2882: DtypeWarn

Columns (2,5,21,34) have mixed types.Specify dtype option on import or set low_memory=F

/usr/local/lib/python3.7/dist-packages/IPython/core/interactiveshell.py:2882: DtypeWarn

Columns (2,5) have mixed types.Specify dtype option on import or set low_memory=False.
```

```
##  Se concatenan los dataframe del 2017 al 2020 'Titulados Educación superior'  ##
action_concat = [df_2017, df_2018, df_2019, df_2020]
df_allYears = pd.concat(action_concat)


## Información general de las columnas del dataset concatenado ##
df_allYears.info()
```

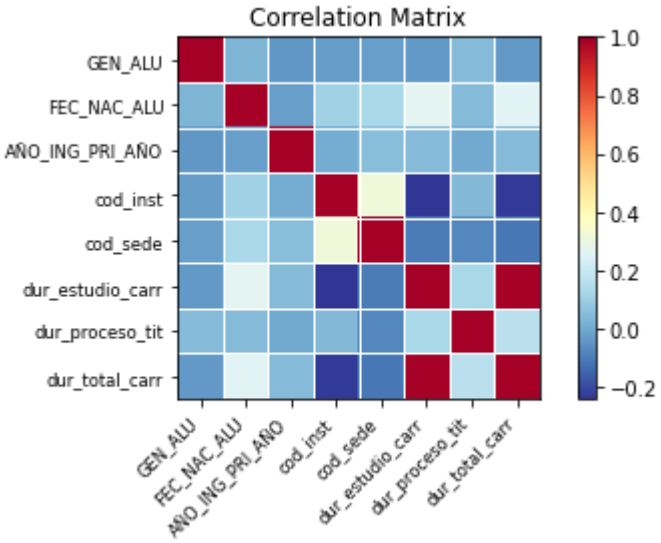
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 935597 entries, 0 to 199033
Data columns (total 40 columns):
 #   Column                   Non-Null Count   Dtype 
---  ------                   --------------   ----- 
 0   cat_periodo              935597 non-null  object
 1   codigo_unico             935597 non-null  object
 2   MRUN                     935597 non-null  object
 3   GEN_ALU                  935597 non-null  int64 
 4   FEC_NAC_ALU              935597 non-null  int64 
 5   EDAD_ALU                 935597 non-null  object
 6   rango_edad               935597 non-null  object
 7   AÑO_ING_PRI_AÑO          935597 non-null  int64 
 8   SEM_ING_PRI_AÑO          935597 non-null  object
 9   AÑO_ING_CARR             935597 non-null  object
 10  SEM_ING_CARR             935597 non-null  object
 11  nomb_titulo_obtenido     935597 non-null  object
 12  nomb_grado_obtenido      935597 non-null  object
 13  FECHA_OBTENCION_TITULO   935597 non-null  int64 
```

```
 14  tipo_inst_1              935597 non-null  object
 15  tipo_inst_2              935597 non-null  object
 16  tipo_inst_3              935597 non-null  object
 17  cod_inst                 935597 non-null  int64
 18  nomb_inst                935597 non-null  object
 19  cod_sede                 935597 non-null  int64
 20  nomb_sede                935597 non-null  object
 21  cod_carrera              935597 non-null  object
 22  nomb_carrera             935597 non-null  object
 23  nivel_global             935597 non-null  object
 24  nivel_carrera_1          935597 non-null  object
 25  nivel_carrera_2          935597 non-null  object
 26  dur_estudio_carr         935597 non-null  int64
 27  dur_proceso_tit          935597 non-null  int64
 28  dur_total_carr           935597 non-null  int64
 29  region_sede              935597 non-null  object
 30  provincia_sede           935597 non-null  object
 31  comuna_sede              935597 non-null  object
 32  jornada                  935597 non-null  object
 33  modalidad                935597 non-null  object
 34  version                  935597 non-null  object
 35  tipo_plan_carr           935597 non-null  object
 36  AREA_CINEUNESCO          935597 non-null  object
 37  oecd_area                935597 non-null  object
 38  oecd_subarea             935597 non-null  object
 39  AREA_CARRERA_GENERICA_N  935597 non-null  object
dtypes: int64(9), object(31)
memory usage: 292.7+ MB
```

```
## Mostrar 5 primeras filas del dataset concatenado ##

df_allYears.head()
```

	cat_periodo	codigo_unico	MRUN	GEN_ALU	FEC_NAC_ALU	EDAD_ALU	rango_edad	AÑ
0	TIT_2017	I100S10C10J1V1	5073113	2	198304	34	30 A 34 AÑOS	

```
## Matriz de correlación entre las columnas dentro del dataset concatenado con los años 2017

corr = df_allYears.set_index('FECHA_OBTENCION_TITULO').corr()
sm.graphics.plot_corr(corr, xnames=list(corr.columns))
plt.figure(figsize=(8,8))
plt.show()
```




```
<Figure size 576x576 with 0 Axes>
```

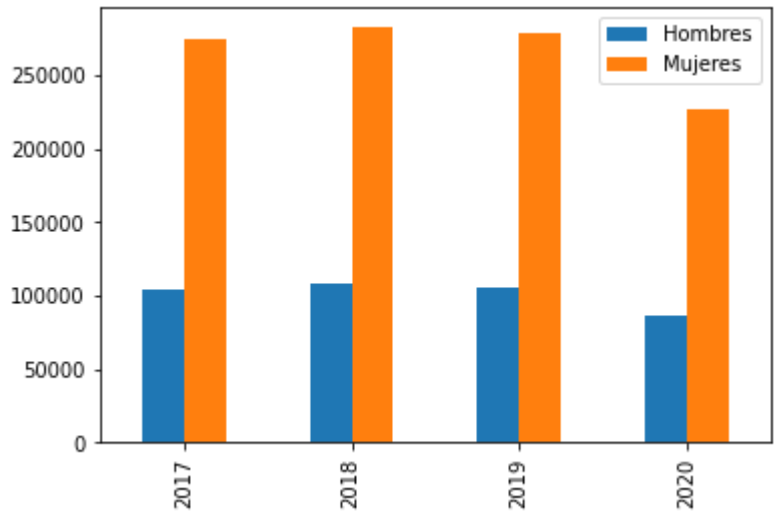
```
## Limpieza de los años ##

df_allYears['cat_periodo'] = df_allYears['cat_periodo'].str.strip("TIT_").astype(int)
df_allYears
```

	cat_periodo	codigo_unico	MRUN	GEN_ALU	FEC_NAC_ALU	EDAD_ALU	rango_ed
0	2017	I100S10C10J1V1	5073113	2	198304	34	30 A AÑ
1	2017	I100S10C10J1V1	5428405	1	198211	34	30 A AÑ
2	2017	I100S10C10J1V1	6006854	2	199410	22	20 A AÑ
3	2017	I100S10C10J1V1	6553554	2	199101	26	25 A AÑ
4	2017	I100S10C10J1V1	7147996	2	199310	23	20 A AÑ
...	...	...	...	...	...	...	
199029	2020	I9S9C7J2V1	18517171	2	199404	26	25 A AÑ
199030	2020	I9S9C7J2V1	20926782	2	199108	28	25 A AÑ
199031	2020	I9S9C7J2V1	22309258	2	199411	25	25 A AÑ

```
## Se recatan las columnas con los datos mas relevantes a analizar ##

df_bd = df_bd[["Año","Nombre Institución","Tipo Institución","Nombre de la Sede","Nombre Regi
               "Area Conocimiento","Tipo Carrera","Nombre del Campus","Duración (en semestres
               "Matrícula primer año hombres","Matrícula primer año mujeres","Matrícula prime
               "Matrícula total hombres","Matrícula total mujeres","Matrícula total extranjer



## Agrupamos por año a los hombres y mujeres en un dataframe por separado ##
df_men = df_allYears[df_allYears["GEN_ALU"] == 1].groupby(by = 'cat_periodo')['GEN_ALU'].sum(
df_women= df_allYears[df_allYears["GEN_ALU"] == 2].groupby(by = 'cat_periodo')['GEN_ALU'].sum
```

```
## Generamos un array con enteros unicos  (años)##
anios = df_allYears['cat_periodo'].unique()



## Hombres y mujeres titulados de educación superior entre los años 2017 al 2020 ##

df_plot = pd.DataFrame({'Hombres': df_men,
                        'Mujeres': df_women},
                       index=anios)

df_plot.plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f497539c8d0>
```




```
 ### Cantidad de matriculas totales en los años 2017 al 2021 ###

tot_matricu = df_bd.groupby(["Año"], sort = False)["Matrícula Total"].sum()

plt_tot_matricu = pd.DataFrame(tot_matricu)

plt_tot_matricu.plot(kind='bar', title = 'Cantidad de matriculas totales en los años 2017 al
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f49752f2290>
```


```
df_bd_institution = df_bd.groupby(["Tipo Institución"], sort = False)[["Matrícula total hombr
"Matrícula total mujeres","Matrícula total extranjeros","Matrícula Total"]].sum()

df_bd_institution
```

Tipo Institución	Matrícula total hombres	Matrícula total mujeres	Matrícula total extranjeros	Matrícu Tot
Univ.	1651589.0	1964878.0	54790.0	36164
I.P.	923013.0	949914.0	30955.0	18729
C.F.T.	305340.0	357114.0	11274.0	6624

```
plt_institution = df_bd_institution.plot(kind = 'line', figsize = (8,8), xlabel = 'Tipo de ir
ylabel = 'Cantidad de matriculas', title = 'Cantidad total de matriculas por tipo de institu
```

Cantidad total de matriculas por tipo de institución

```
# df_age_2017 = df_allYears.groupby(['cat_periodo']).EDAD_ALU.mean()

# df_age_2017
```

```
## Mostrando los primeros 5 registros del csv obtenido de una base de datos del ministerio ##
df_bd.head()
```

	Año	Nombre Institución	Tipo Institución	Nombre de la Sede	Nombre Region	Mención o Especialidad	Tipo Programa	Conoci
0	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	NaN	Programa Regular	Admini y C
1	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	NaN	Programa Regular	Admini y C
2	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	Ciencias Económicas o Ciencias de la Administr...	Programa Regular	Admini y C
3	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	NaN	Programa Regular	Admini y C
4	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	NaN	Programa Regular	Agr Silv Ve

5 rows × 22 columns

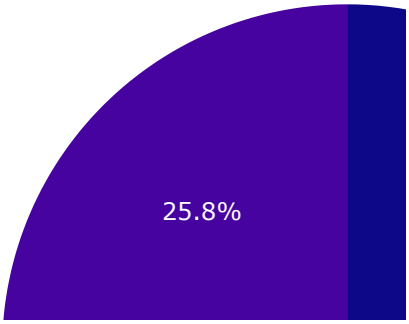
```
df_bd = df_bd.rename(columns={'Valor de arancel':'Valor_arancel'})
df_bd.head()
```

	Año	Nombre Institución	Tipo Institución	Nombre de la Sede	Nombre Region	Mención o Especialidad	Tipo Programa	Conoci
0	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	NaN	Programa Regular	Admini y C
1	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	NaN	Programa Regular	Admini y C
2	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	Ciencias Económicas o Ciencias de la Administr...	Programa Regular	Admini y C
3	2021	U. DE CHILE	Univ.	Santiago	Región Metropolitana	NaN	Programa Regular	Admini y C

Agr

```
## Renombramos las columnas VMATRICULA, VARANCEL y VTITULO ##
df_bd.rename(columns={'Valor de matrícula':'VMATRICULA', 'Valor de arancel': 'VARANCEL', 'Val


## Promedio del valor de la matricula por tipo de institución ##

torta_0 = df_bd.groupby(['Tipo Institución']).VMATRICULA.mean()
torta = pd.DataFrame(torta_0)
torta.reset_index(drop=False, inplace=True)

fig = px.pie(torta, values='VMATRICULA',
             names='Tipo Institución',
             title = 'Promedio valor de matrícula por tipo de institución',
             color_discrete_sequence = px.colors.sequential.Plasma)
fig.show()
```

Promedio valor de matrícula por tipo de institución




```
df_2018.head()
```

	cat_periodo	codigo_unico	MRUN	GEN_ALU	FEC_NAC_ALU	EDAD_ALU	rango_edad	AÑ
0	TIT_2018	I100S10C10J1V1	590763	2	199606	22	20 A 24 AÑOS	
1	TIT_2018	I100S10C10J1V1	1979973	2	198510	32	30 A 34 AÑOS	
2	TIT_2018	I100S10C10J1V1	2988701	1	199510	22	20 A 24 AÑOS	
3	TIT_2018	I100S10C10J1V1	7059325	1	199509	22	20 A 24 AÑOS	
4	TIT_2018	I100S10C10J1V1	8877808	2	199603	22	20 A 24 AÑOS	

5 rows × 40 columns

```
df_2018.describe()
```

	GEN_ALU	FEC_NAC_ALU	AÑO_ING_PRI_AÑO	SEM_ING_PRI_AÑO	AÑO_ING_CARR	S
count	249290.000000	249290.000000	249290.000000	249290.000000	249290.000000	24
mean	1.566300	198882.615737	2013.767740	1.050054	2014.633102	
std	0.495586	766.722418	8.004359	0.237415	2.476645	
min	1.000000	190001.000000	1900.000000	0.000000	1970.000000	
25%	1.000000	198603.000000	2013.000000	1.000000	2013.000000	
50%	2.000000	199109.000000	2015.000000	1.000000	2015.000000	

```
df_2018.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 249290 entries, 0 to 249289
Data columns (total 40 columns):
 #   Column                   Non-Null Count   Dtype 
---  ------                   --------------   ----- 
 0   cat_periodo              249290 non-null  object
 1   codigo_unico             249290 non-null  object
 2   MRUN                     249290 non-null  object
 3   GEN_ALU                  249290 non-null  int64 
 4   FEC_NAC_ALU              249290 non-null  int64 
 5   EDAD_ALU                 249290 non-null  object
 6   rango_edad               249290 non-null  object
 7   AÑO_ING_PRI_AÑO          249290 non-null  int64 
 8   SEM_ING_PRI_AÑO          249290 non-null  int64 
 9   AÑO_ING_CARR             249290 non-null  int64 
 10  SEM_ING_CARR             249290 non-null  int64 
 11  nomb_titulo_obtenido     249290 non-null  object
 12  nomb_grado_obtenido      249290 non-null  object
 13  FECHA_OBTENCION_TITULO   249290 non-null  int64 
 14  tipo_inst_1              249290 non-null  object
 15  tipo_inst_2              249290 non-null  object
 16  tipo_inst_3              249290 non-null  object
 17  cod_inst                 249290 non-null  int64 
 18  nomb_inst                249290 non-null  object
 19  cod_sede                 249290 non-null  int64 
 20  nomb_sede                249290 non-null  object
 21  cod_carrera              249290 non-null  object
 22  nomb_carrera             249290 non-null  object
 23  nivel_global             249290 non-null  object
 24  nivel_carrera_1          249290 non-null  object
 25  nivel_carrera_2          249290 non-null  object
 26  dur_estudio_carr         249290 non-null  int64 
 27  dur_proceso_tit          249290 non-null  int64 
 28  dur_total_carr           249290 non-null  int64 
 29  region_sede              249290 non-null  object
 30  provincia_sede           249290 non-null  object
 31  comuna_sede              249290 non-null  object
 32  jornada                  249290 non-null  object
 33  modalidad                249290 non-null  object
 34  version                  249290 non-null  object
```

```
 35  tipo_plan_carr           249290 non-null  object
 36  AREA_CINEUNESCO          249290 non-null  object
 37  oecd_area                249290 non-null  object
 38  oecd_subarea             249290 non-null  object
 39  AREA_CARRERA_GENERICA_N  249290 non-null  object
dtypes: int64(12), object(28)
memory usage: 76.1+ MB
```

```
edad_alumno = pd.to_numeric(df_2018.EDAD_ALU, errors='coerce')
```

```
edad_alumno.min()
```

```
18.0
```

```
edad_alumno.max()
```

```
80.0
```

```
edad_alumno.mean()
```

```
28.719100131979573
```

```
##  Se cambia el nombre de la columna GEN_ALU  ##

df_2020.rename(columns={'GEN_ALU':'GENERO'}, inplace=True)
df_2019.rename(columns={'GEN_ALU':'GENERO'}, inplace=True)
df_2018.rename(columns={'GEN_ALU':'GENERO'}, inplace=True)
df_2017.rename(columns={'GEN_ALU':'GENERO'}, inplace=True)
```

```
##  Se cruzan los hombres y mujeres con el tipo de jornada y se agrupan en una tabla ##

prueba_2020 = pd.crosstab(df_2020.GENERO, df_2020.jornada)
```

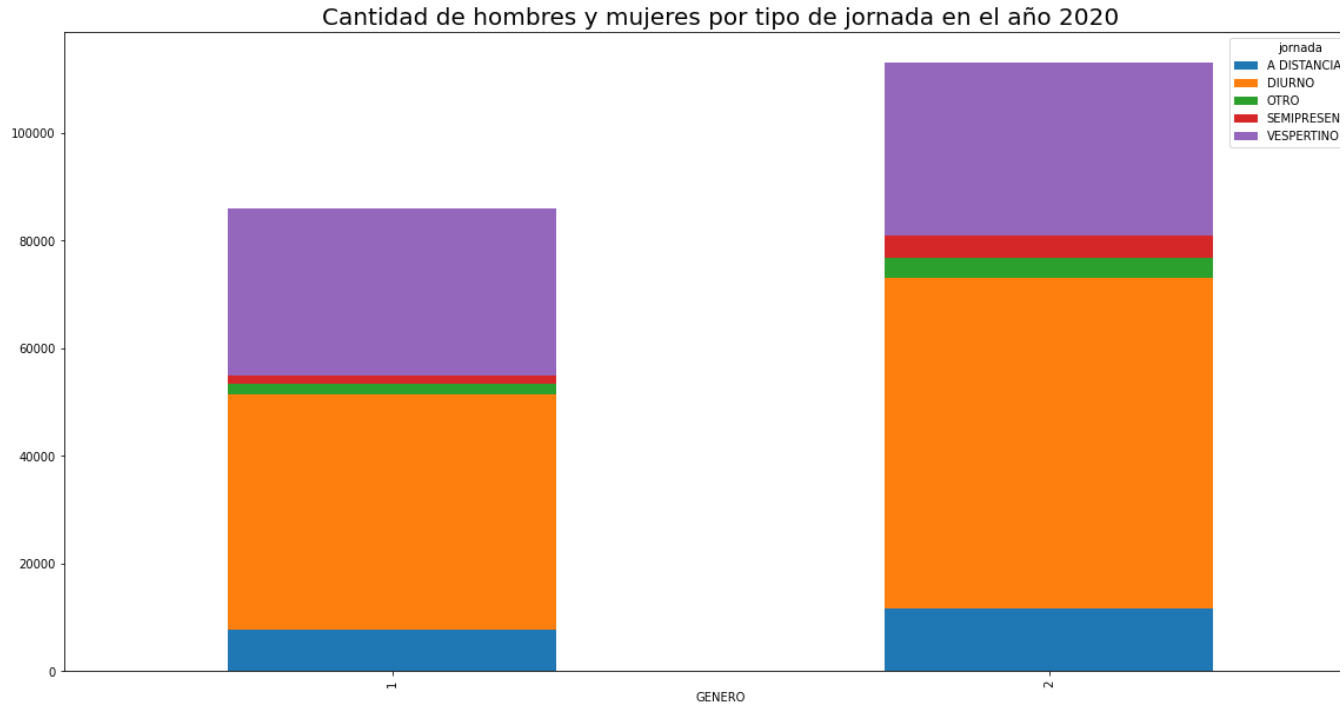
```
prueba_2019 = pd.crosstab(df_2019.GENERO, df_2019.jornada)
```

```
prueba_2018 = pd.crosstab(df_2018.GENERO, df_2018.jornada)
```

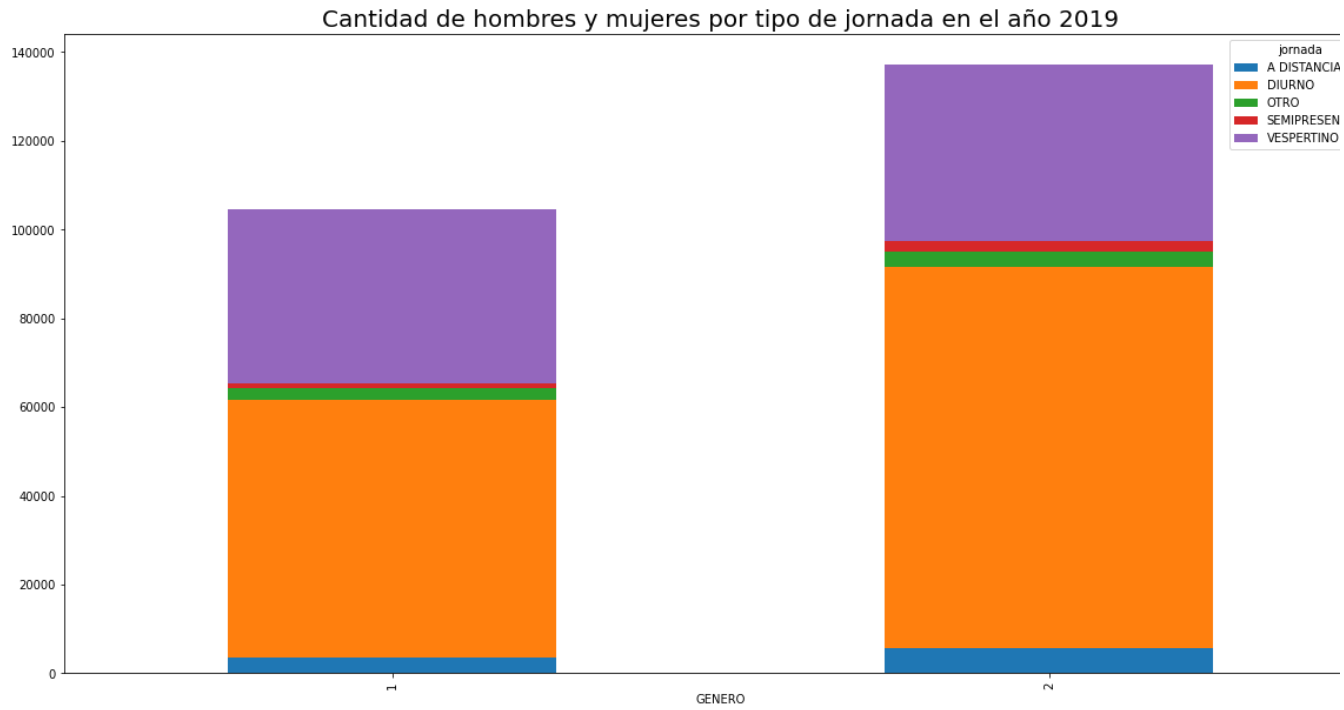
```
prueba_2017 = pd.crosstab(df_2017.GENERO, df_2017.jornada)
```

```
##  Cantidad de hombres vs mujeres por el tipo de jornada en el año 2020 ##
```

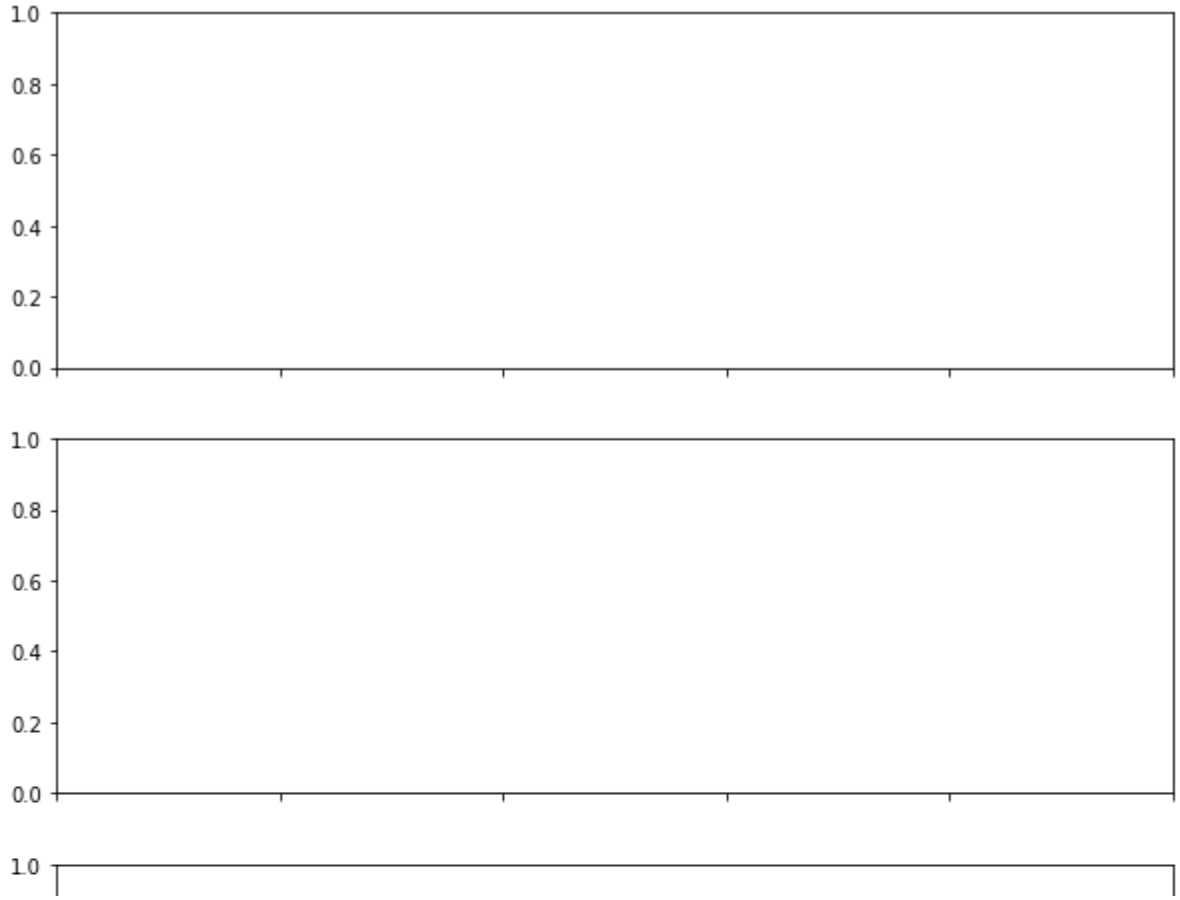
```
apilado = prueba_2020.plot(kind="bar", stacked=True, figsize=(20,10))
titulo_3 = "Cantidad de hombres y mujeres por tipo de jornada en el año 2020"
plt.title(titulo_3, fontsize=20)
plt.show()
```




```
apilado2 = prueba_2017.plot(kind="bar", stacked=True, figsize=(20,10))
titulo_4 = "Cantidad de hombres y mujeres por tipo de jornada en el año 2019"
plt.title(titulo_4, fontsize=20)
plt.show()
```




```
fig, axes = plt.subplots(nrows=4, ncols=1, figsize = (10,15), sharex = True, sharey = True) #
```




```
fig, axes = plt.subplots(nrows=1, ncols=3,figsize = (25,10), sharex = True, sharey = True) #s
```


```
df_allYears.head(3)
```

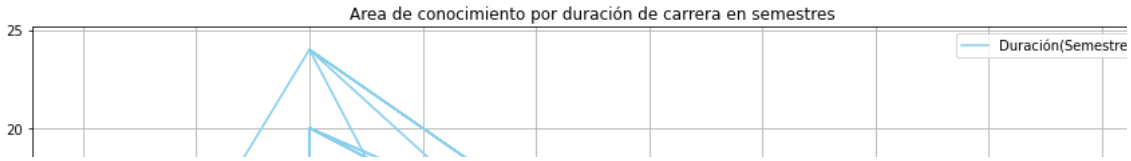
	cat_periodo	codigo_unico	MRUN	GEN_ALU	FEC_NAC_ALU	EDAD_ALU	rango_edad	AÑ
0	2017	I100S10C10J1V1	5073113	2	198304	34	30 A 34 AÑOS	
1	2017	I100S10C10J1V1	5428405	1	198211	34	30 A 34 AÑOS	
2	2017	I100S10C10J1V1	6006854	2	199410	22	20 A 24 AÑOS	

3 rows × 40 columns

```
df_allYears['EDAD_ALU'] = pd.to_numeric(df_allYears['EDAD_ALU'],errors='coerce')

df_allYears["EDAD_ALU"] = df_allYears["EDAD_ALU"].fillna(df_allYears["EDAD_ALU"].mean())


df_allYears['EDAD_ALU'] = df_allYears['EDAD_ALU'].astype(int)

fig, ax = plt.subplots(nrows=2,ncols=1,figsize=(15,15))


ax[0].plot(df_allYears['AREA_CINEUNESCO'], df_allYears['dur_estudio_carr'] , color='skyblue',
ax[0].grid(True)
ax[0].legend()
ax[0].set_xlabel('Duración de la carrera (Semestres)')
ax[0].set_ylabel('Area de conocimiento')
ax[0].set_title('Area de conocimiento por duración de carrera en semestres')


ax[1].plot(df_allYears['EDAD_ALU'], df_allYears['AREA_CINEUNESCO'], color='#444444', linesty
ax[1].grid(True)
```

```
    ax[1].legend()
    ax[1].set_xlabel('Edade del alumno')
    ax[1].set_ylabel('Area de conocimiento')
    ax[1].set_title('Edad del alumno por area de conocimiento')
```

```
Text(0.5, 1.0, 'Edad del alumno por area de conocimiento')
```




```
#Creamos archivo .db en el directorio del programa
connection = sqlite3.connect('PostgreSQL.db')
c = connection.cursor()

#Creamos la tabla donde se cargará el dataframe
c.execute('''
          CREATE TABLE IF NOT EXISTS graduates
          (
            [cat_periodo] INTEGER,
            [codigo_unico] INTEGER PRIMARY KEY,
            [MRUN] INTEGER,
            [GEN_ALU] INTEGER,
            [FEC_NAC_ALU[ TEXT,
            [EDAD_ALU] INTEGER,
            [rango_edad] TXT,
            [AÑO_ING_PRI_AÑO] INTEGER,
            [SEM_ING_PRI_AÑO] INTEGER,
            [AÑO_ING_CARR] INTEGER,
            [SEM_ING_CARR] INTEGER,
            [nomb_titulo_obtenido] VARCHAR,
            [nomb_grado_obtenido] VARCHAR,
            [FECHA_OBTENCION_TITULO] VARCHAR,
            [tipo_inst_1] VARCHAR,
            [tipo_inst_2] VARCHAR,
            [tipo_inst_3] VARCHAR,
            [cod_inst] INTEGER,
            [nomb_inst] ,
            [cod_sede],
            [nomb_sede] VARCHAR,
            [cod_carrera],
            [nomb_carrera],
            [nivel_global],
            [nivel_carrera_1],
            [nivel_carrera_2],
            [dur_estudio_carr],
            [dur_proceso_tit],
            [dur_total_carr],
            [region_sede] VARCHAR,
            [provincia_sede] VARCHAR,
            [comuna_sede] VARCHAR,
            [jornada] VARCHAR,
            [modalidad] VARCHAR,
            [version] INTEGER,
            [tipo_plan_carr] VARCHAR,
            [AREA_CINEUNESCO] VARCHAR,
            [oecd_area] VARCHAR
```

```
            [oecd_area] VARCHAR,
            [oecd_subarea] VARCHAR,
            [AREA_CARRERA_GENERICA_N] VARCHAR)''')

connection.commit()
#Ejemplo de dataframe creado en pandas en base a una colección del tipo diccionario

df_allYears.to_sql('graduates', connection, if_exists='replace', index = False)
#Consultamos la tabla creada en la base de datos .db para ver que proceso fue exitoso
df = pd.read_sql('SELECT * FROM graduates', connection)
print(df)
```

```
     935593      DIGUILLIN       CHILLAN  VESPERTINO  PRESENCIAL       1
     935594      DIGUILLIN       CHILLAN  VESPERTINO  PRESENCIAL       1
     935595      DIGUILLIN       CHILLAN  VESPERTINO  PRESENCIAL       1
     935596      DIGUILLIN       CHILLAN  VESPERTINO  PRESENCIAL       1

            tipo_plan_carr    AREA_CINEUNESCO  \
     0        PLAN REGULAR         TECNOLOGÍA
     1        PLAN REGULAR         TECNOLOGÍA
     2        PLAN REGULAR         TECNOLOGÍA
     3        PLAN REGULAR         TECNOLOGÍA
     4        PLAN REGULAR         TECNOLOGÍA
     ...               ...                ...
     935592   PLAN REGULAR  CIENCIAS SOCIALES
     935593   PLAN REGULAR  CIENCIAS SOCIALES
     935594   PLAN REGULAR  CIENCIAS SOCIALES
     935595   PLAN REGULAR  CIENCIAS SOCIALES
     935596   PLAN REGULAR  CIENCIAS SOCIALES

                                                 oecd_area  \
     0                                           SERVICIOS
     1                                           SERVICIOS
     2                                           SERVICIOS
     3                                           SERVICIOS
     4                                           SERVICIOS
     ...                                               ...
     935592  CIENCIAS SOCIALES, ENSEÑANZA COMERCIAL Y DERECHO
     935593  CIENCIAS SOCIALES, ENSEÑANZA COMERCIAL Y DERECHO
     935594  CIENCIAS SOCIALES, ENSEÑANZA COMERCIAL Y DERECHO
     935595  CIENCIAS SOCIALES, ENSEÑANZA COMERCIAL Y DERECHO
     935596  CIENCIAS SOCIALES, ENSEÑANZA COMERCIAL Y DERECHO

                                      oecd_subarea  \
     0                       SERVICIOS DE SEGURIDAD
     1                       SERVICIOS DE SEGURIDAD
     2                       SERVICIOS DE SEGURIDAD
     3                       SERVICIOS DE SEGURIDAD
     4                       SERVICIOS DE SEGURIDAD
     ...                                        ...
     935592  CIENCIAS SOCIALES Y DEL COMPORTAMIENTO
     935593  CIENCIAS SOCIALES Y DEL COMPORTAMIENTO
     935594  CIENCIAS SOCIALES Y DEL COMPORTAMIENTO
     935595  CIENCIAS SOCIALES Y DEL COMPORTAMIENTO
     935596  CIENCIAS SOCIALES Y DEL COMPORTAMIENTO

                        AREA_CARRERA_GENERICA_N
```

```
                    AREA_CARRERA_GENERICA_N
0       INGENIERÍA EN PREVENCIÓN DE RIESGOS
1       INGENIERÍA EN PREVENCIÓN DE RIESGOS
2       INGENIERÍA EN PREVENCIÓN DE RIESGOS
3       INGENIERÍA EN PREVENCIÓN DE RIESGOS
4       INGENIERÍA EN PREVENCIÓN DE RIESGOS
...                                     ...
935592                          PSICOLOGÍA
935593                          PSICOLOGÍA
935594                          PSICOLOGÍA
935595                          PSICOLOGÍA
935596                          PSICOLOGÍA

[935597 rows x 40 columns]
```

✓ 15 s completado a las 18:52