PROJECT INTRODUCTION

A wage is that the distribution from associate degree leader of a security (expected come back or profits derived alone from others) paid to associate degree worker. Like interest is paid resolute associate degree capitalist on his investments, a wage is paid as earnings to the worker on his endowed assets (time, money, labor, resources, and thought). Some samples of wage distributions embrace counteractive payments like earnings, prevailing wage, and yearly bonuses, and remunerative payments like prizes and tip payouts.

Wages area unit a part of the expenses that area unit concerned in running a business, and add price to the worker in honor of his principal protected note or internet investment.

Payment by wage contrasts with salaried work, during which the leader pays associate degree organized quantity at steady intervals (such as every week or month) in spite of hours worked, with commission that conditions pay on individual performance, and with compensation supported the performance of the corporate as a full. Waged workers can also receive tips or gratuity paid directly by purchasers and worker edges that area unit non-monetary styles of compensation. Since wage labour is that the predominant type of work, the term "wage" typically refers to any or all forms (or all financial forms) of worker compensation.

An hourly employee or hourly worker is associate degree worker paid associate degree hourly wage for his or her services, as opposition a set remuneration. Hourly staff could typically be found in commission and producing occupations, however area unit common across a range of fields. Hourly employment is commonly associated however not synonymous with at-will employment. As of Sep 2017, the earnings within the u. s. for hourly staff is $7.25 per hour or $2.13 per hour for a tipped worker. As a tipped worker, wages and tips should equal the quality earnings or the leader is needed to produce the distinction.

ANALYSIS OBJECTIVES

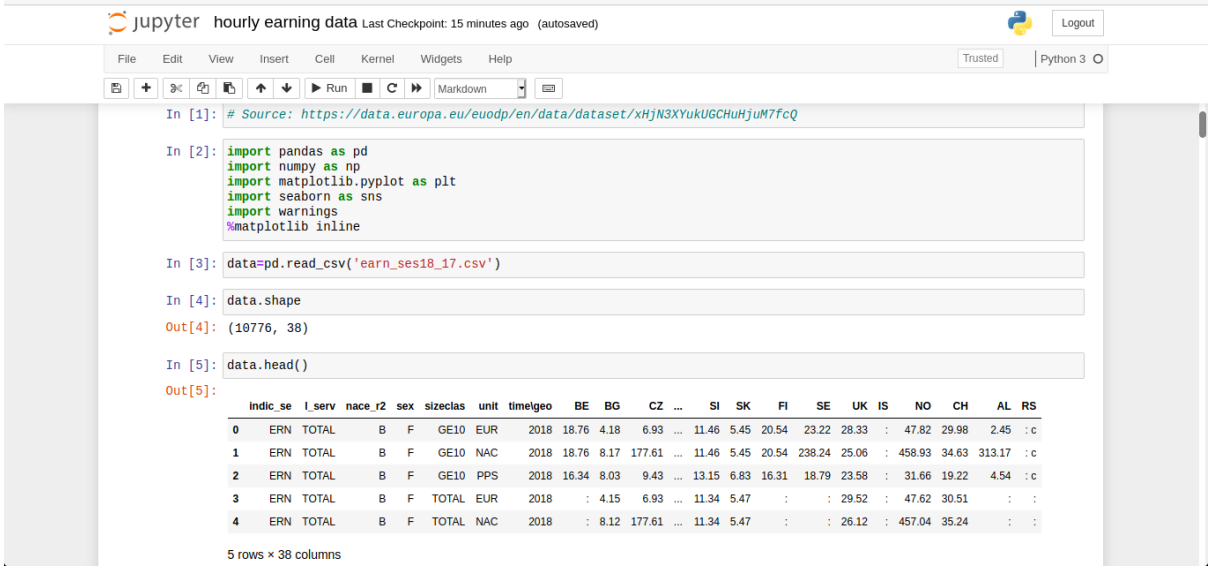
Here are the list of question whose analysis which be given below..

1. Which are the different countries from where data has been collected
2. What is the male to female to transgender (no. of entries)
3. Max, mean and median income in 10 country for all different genders
4. Average income in top 10 countries irrespective of the gender
5. What is the different types of sizeclas and their counts
6. What is the different units of currencies used to pay and their counts
7. Which country pays the maximum average salary and by how much.

DATA ACQUISITION AND CLEANING

Code to read the data from Excel / CSV / HTML.

To read the dataset in csv format, we will load it into Pandas data frame but first let's import the pandas library and set an alias by typing **"import pandas as pd"**. After importing the library with the alias **"pd"**, let us load the .csv file using the following line of code:



jupyter hourly earning data Last Checkpoint: 15 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Markdown

```
In [1]: # Source: https://data.europa.eu/euodp/en/data/dataset/xHjN3XYukUGCHuHjuM7fcQ

In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        import warnings
        %matplotlib inline

In [3]: data=pd.read_csv('earn_ses18_17.csv')

In [4]: data.shape
Out[4]: (10776, 38)

In [5]: data.head()
Out[5]:
```

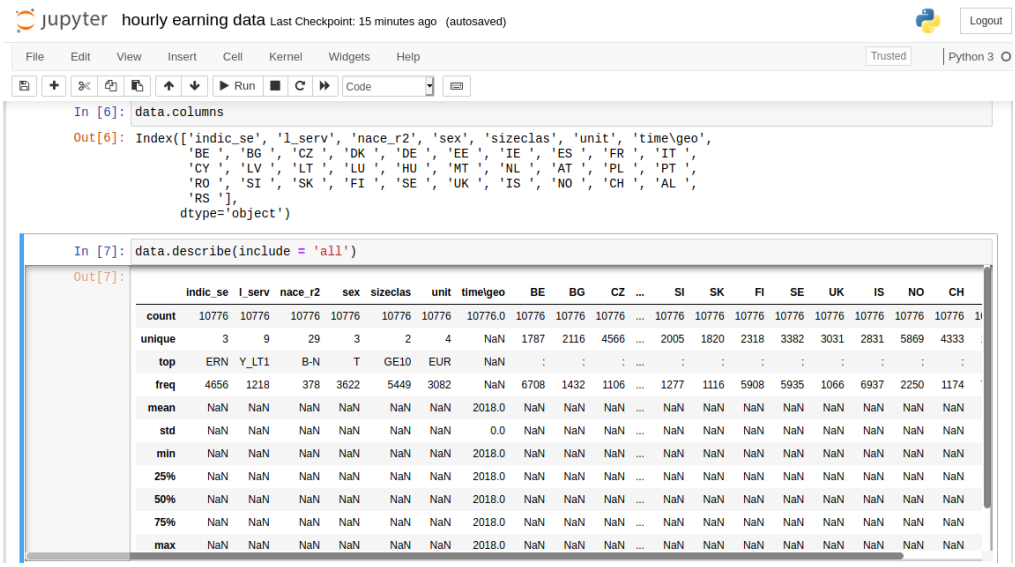
	indic_se	l_serv	nace_r2	sex	sizeclas	unit	time\geo	BE	BG	CZ	...	SI	SK	FI	SE	UK	IS	NO	CH	AL	RS
0	ERN	TOTAL	B	F	GE10	EUR	2018	18.76	4.18	6.93	...	11.46	5.45	20.54	23.22	28.33	:	47.82	29.98	2.45	: c
1	ERN	TOTAL	B	F	GE10	NAC	2018	18.76	8.17	177.61	...	11.46	5.45	20.54	238.24	25.06	:	458.93	34.63	313.17	: c
2	ERN	TOTAL	B	F	GE10	PPS	2018	16.34	8.03	9.43	...	13.15	6.83	16.31	18.79	23.58	:	31.66	19.22	4.54	: c
3	ERN	TOTAL	B	F	TOTAL	EUR	2018	:	4.15	6.93	...	11.34	5.47	:	:	29.52	:	47.62	30.51	:	:
4	ERN	TOTAL	B	F	TOTAL	NAC	2018	:	8.12	177.61	...	11.34	5.47	:	:	26.12	:	457.04	35.24	:	:

5 rows × 38 columns

Here we have import our csv files and read through pandas library.

Here the csv file can be read through (Pandas library) and store in two different dataframes. The Dataframe can be shown through **.head()**. The number of rows we want to show, that number we have to pass in head parentheses as an argument.

Now if we want to describe our dataframe for our better understanding to know the stats. and other parameter that our dataset should follow



jupyter hourly earning data Last Checkpoint: 15 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

```
In [6]: data.columns
Out[6]: Index(['indic_se', 'l_serv', 'nace_r2', 'sex', 'sizeclas', 'unit', 'time\geo',
               'BE ', 'BG ', 'CZ ', 'DK ', 'DE ', 'EE ', 'IE ', 'ES ', 'FR ', 'IT ',
               'CY ', 'LV ', 'LT ', 'LU ', 'HU ', 'MT ', 'NL ', 'AT ', 'PL ', 'PT ',
               'RO ', 'SI ', 'SK ', 'FI ', 'SE ', 'UK ', 'IS ', 'NO ', 'CH ', 'AL ',
               'RS '],
              dtype='object')

In [7]: data.describe(include = 'all')
Out[7]:
```

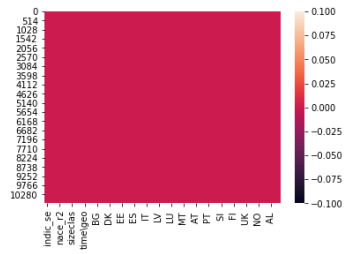
	indic_se	l_serv	nace_r2	sex	sizeclas	unit	time\geo	BE	BG	CZ	...	SI	SK	FI	SE	UK	IS	NO	CH
count	10776	10776	10776	10776	10776	10776	10776.0	10776	10776	10776	...	10776	10776	10776	10776	10776	10776	10776	10776
unique	3	9	29	3	2	4	NaN	1787	2116	4566	...	2005	1820	2318	3382	3031	2831	5869	4333
top	ERN	Y_LT1	B-N	T	GE10	EUR	NaN	:	:	:	...	:	:	:	:	:	:	:	:
freq	4656	1218	378	3622	5449	3082	NaN	6708	1432	1106	...	1277	1116	5908	5935	1066	6937	2250	1174
mean	NaN	NaN	NaN	NaN	NaN	NaN	2018.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	2018.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	2018.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	2018.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	2018.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	2018.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Clean the unnecessary data, by removing, replace the missing data and renaming the columns.

Dataset generally contains some null value, which is generally caused by misplacing some values. So its necessary to clean this mess from our dataset for better visualization

In the process of data cleaning we have to drop the columns which contains atmost null value, but there is no column. Then subsequently we have to fill the other null values column with its mean value, so that the values will not much more effect. Finally after all process our heatmap diagram is pink which shows that our datasets is almost clean.

```
In [11]: data = data.replace(': ',np.NaN)
         data = data.replace(': c',np.NaN)
         for x in data.columns[7:]:
             data[x] = data[x].astype(float)
         data = data.fillna(data.mean())

In [12]: sns.heatmap(data.isnull())
Out[12]: <AxesSubplot:>
```



why data clean needed (for your data)

Data cleansing or scrubbing or appending is the procedure of correcting or removing inaccurate and corrupt data. This process is crucial and emphasized because wrong data can drive a business to wrong decisions, conclusions, and poor analysis, especially if the huge quantities of big data are into the picture.

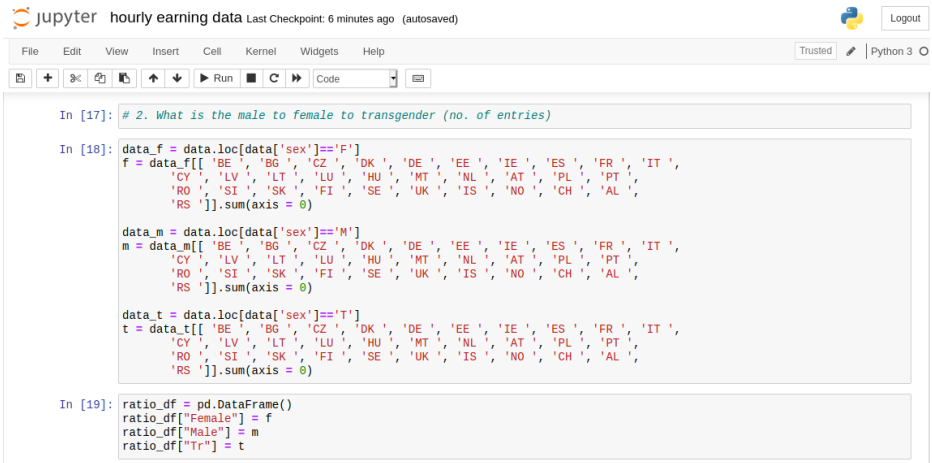
DATA AND EXPLORATORY ANALYSIS

Code and its output with Explanation

1. Which are the different countries from where data has been collected


```
In [13]: # 1. Which are the different countries from where data has been collected

In [14]: countries = data.columns[7:]
         print("Number of countries is :",len(countries),
               "\n Name: ", countries)

         Number of countries is : 31
          Name:  Index(['BE ', 'BG ', 'CZ ', 'DK ', 'DE ', 'EE ', 'IE ', 'ES ', 'FR ', 'IT ',
                'CY ', 'LV ', 'LT ', 'LU ', 'HU ', 'MT ', 'NL ', 'AT ', 'PL ', 'PT ',
                'RO ', 'SI ', 'SK ', 'FI ', 'SE ', 'UK ', 'IS ', 'NO ', 'CH ', 'AL ',
                'RS '],
               dtype='object')

In [15]: df = pd.DataFrame({'Country': countries})
```

2. What is the male to female to transgender (no. of entries)



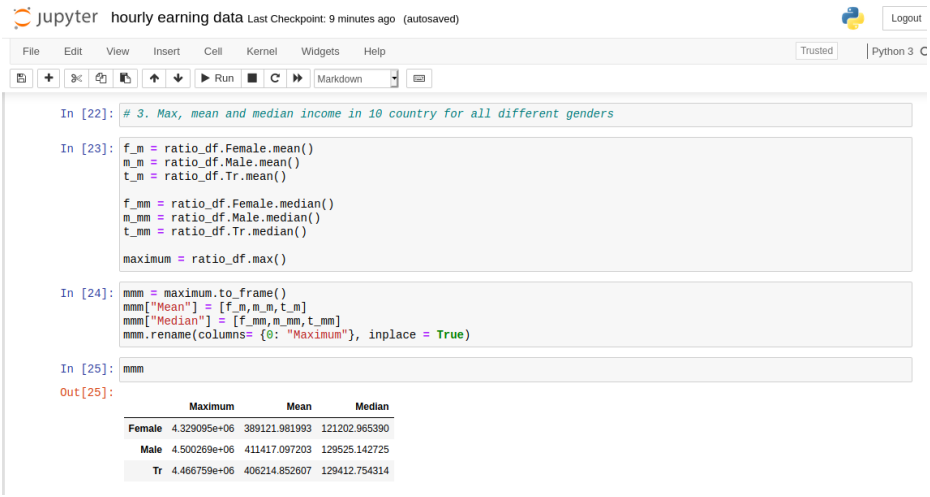

```
In [17]: # 2. What is the male to female to transgender (no. of entries)

In [18]: data_f = data.loc[data['sex']=='F']
         f = data_f[[ 'BE ', 'BG ', 'CZ ', 'DK ', 'DE ', 'EE ', 'IE ', 'ES ', 'FR ', 'IT ',
                'CY ', 'LV ', 'LT ', 'LU ', 'HU ', 'MT ', 'NL ', 'AT ', 'PL ', 'PT ',
                'RO ', 'SI ', 'SK ', 'FI ', 'SE ', 'UK ', 'IS ', 'NO ', 'CH ', 'AL ',
                'RS ']].sum(axis = 0)

         data_m = data.loc[data['sex']=='M']
         m = data_m[[ 'BE ', 'BG ', 'CZ ', 'DK ', 'DE ', 'EE ', 'IE ', 'ES ', 'FR ', 'IT ',
                'CY ', 'LV ', 'LT ', 'LU ', 'HU ', 'MT ', 'NL ', 'AT ', 'PL ', 'PT ',
                'RO ', 'SI ', 'SK ', 'FI ', 'SE ', 'UK ', 'IS ', 'NO ', 'CH ', 'AL ',
                'RS ']].sum(axis = 0)

         data_t = data.loc[data['sex']=='T']
         t = data_t[[ 'BE ', 'BG ', 'CZ ', 'DK ', 'DE ', 'EE ', 'IE ', 'ES ', 'FR ', 'IT ',
                'CY ', 'LV ', 'LT ', 'LU ', 'HU ', 'MT ', 'NL ', 'AT ', 'PL ', 'PT ',
                'RO ', 'SI ', 'SK ', 'FI ', 'SE ', 'UK ', 'IS ', 'NO ', 'CH ', 'AL ',
                'RS ']].sum(axis = 0)

In [19]: ratio_df = pd.DataFrame()
         ratio_df["Female"] = f
         ratio_df["Male"] = m
         ratio_df["Tr"] = t
```

```
In [20]: ratio_df
```

Out[20]:

	Female	Male	Tr
BE	1.373932e+05	1.371208e+05	1.390924e+05
BG	9.075975e+04	9.322660e+04	9.472309e+04
CZ	2.825609e+05	3.191408e+05	3.061772e+05
DK	4.343217e+05	4.568022e+05	4.532588e+05
DE	1.118471e+05	1.239074e+05	1.204264e+05
EE	9.758899e+04	1.038014e+05	1.025275e+05
IE	1.640755e+05	1.707073e+05	1.694492e+05
ES	1.131806e+05	1.208625e+05	1.189052e+05
FR	1.270261e+05	1.295251e+05	1.299083e+05
IT	1.062603e+05	1.078432e+05	1.085175e+05
CY	1.042323e+05	1.074922e+05	1.071280e+05
LV	9.853218e+04	9.990642e+04	1.000083e+05
LT	9.256444e+04	9.242282e+04	9.571473e+04
LU	1.503700e+05	1.532302e+05	1.543440e+05

3. Max, mean and median income in 10 country for all different genders



jupyter hourly earning data Last Checkpoint: 9 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [22]:
```
# 3. Max, mean and median income in 10 country for all different genders
```

In [23]:
```
f_m = ratio_df.Female.mean()
m_m = ratio_df.Male.mean()
t_m = ratio_df.Tr.mean()

f_mm = ratio_df.Female.median()
m_mm = ratio_df.Male.median()
t_mm = ratio_df.Tr.median()

maximum = ratio_df.max()
```

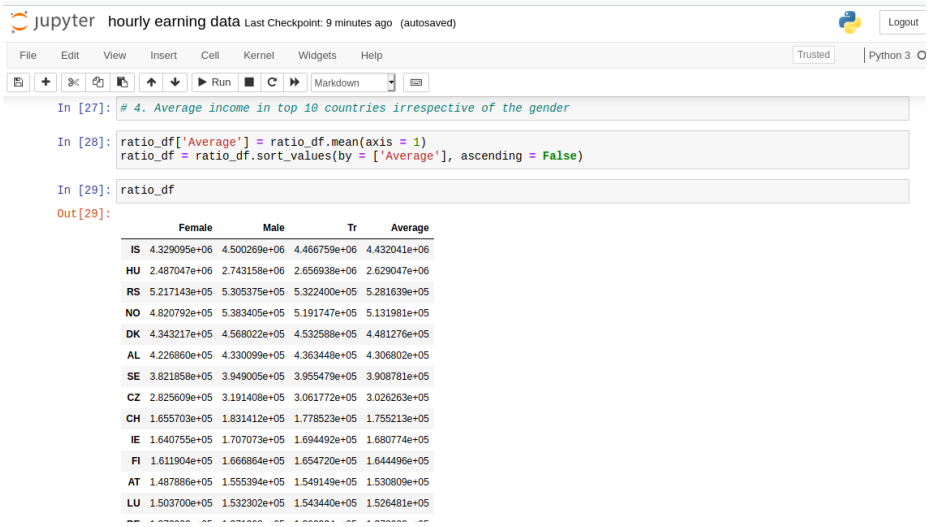
In [24]:
```
mmm = maximum.to_frame()
mmm["Mean"] = [f_m,m_m,t_m]
mmm["Median"] = [f_mm,m_mm,t_mm]
mmm.rename(columns= {0: "Maximum"}, inplace = True)
```

In [25]:
```
mmm
```

Out[25]:

	Maximum	Mean	Median
Female	4.329095e+06	389121.981993	121202.965390
Male	4.500269e+06	411417.097203	129525.142725
Tr	4.466759e+06	406214.852607	129412.754314

4. Average income in top 10 countries irrespective of the gender



jupyter hourly earning data Last Checkpoint: 9 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [27]:
```
# 4. Average income in top 10 countries irrespective of the gender
```

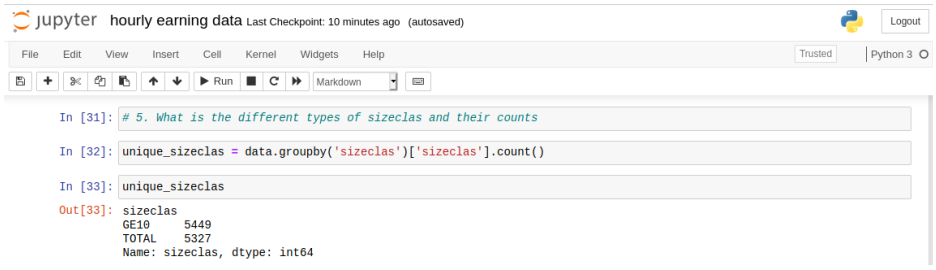
In [28]:
```
ratio_df['Average'] = ratio_df.mean(axis = 1)
ratio_df = ratio_df.sort_values(by = ['Average'], ascending = False)
```

In [29]:
```
ratio_df
```

Out[29]:

	Female	Male	Tr	Average
IS	4.329095e+06	4.500269e+06	4.466759e+06	4.432041e+06
HU	2.487047e+06	2.743158e+06	2.656938e+06	2.629047e+06
RS	5.217143e+05	5.305375e+05	5.322400e+05	5.281639e+05
NO	4.820792e+05	5.383405e+05	5.191747e+05	5.131981e+05
DK	4.343217e+05	4.568022e+05	4.532588e+05	4.481276e+05
AL	4.226860e+05	4.330099e+05	4.363448e+05	4.306802e+05
SE	3.821858e+05	3.949005e+05	3.955479e+05	3.908781e+05
CZ	2.825609e+05	3.191408e+05	3.061772e+05	3.026263e+05
CH	1.655703e+05	1.831412e+05	1.778523e+05	1.755213e+05
IE	1.640755e+05	1.707073e+05	1.694492e+05	1.680774e+05
FI	1.611904e+05	1.666864e+05	1.654720e+05	1.644496e+05
AT	1.487886e+05	1.555394e+05	1.549149e+05	1.530809e+05
LU	1.503700e+05	1.532302e+05	1.543440e+05	1.526481e+05

5. What is the different types of sizeclas and their counts



jupyter hourly earning data Last Checkpoint: 10 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

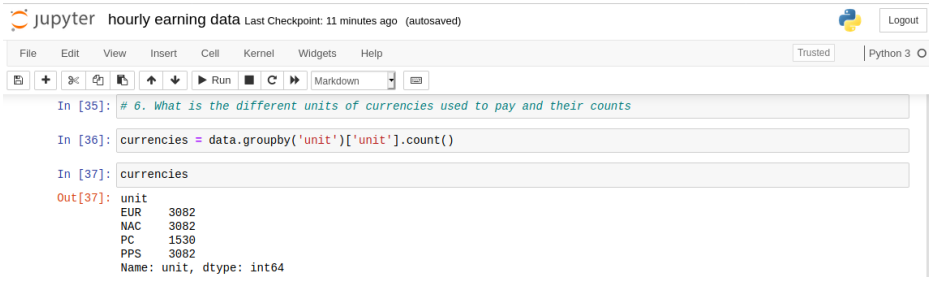
In [31]:
```
# 5. What is the different types of sizeclas and their counts
```

In [32]:
```
unique_sizeclas = data.groupby('sizeclas')['sizeclas'].count()
```

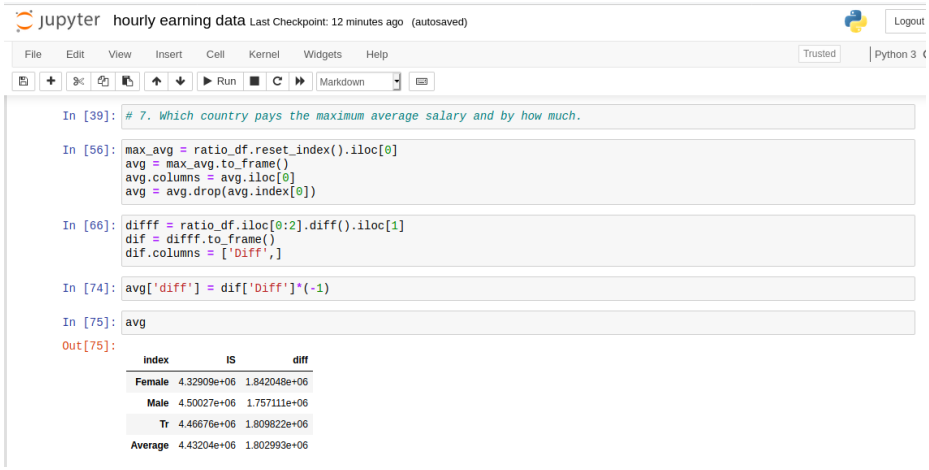
In [33]:
```
unique_sizeclas
```

Out[33]:
```
sizeclas
GE10     5449
TOTAL    5327
Name: sizeclas, dtype: int64
```

6. What is the different units of currencies used to pay and their counts



7. Which country pays the maximum average salary and by how much.

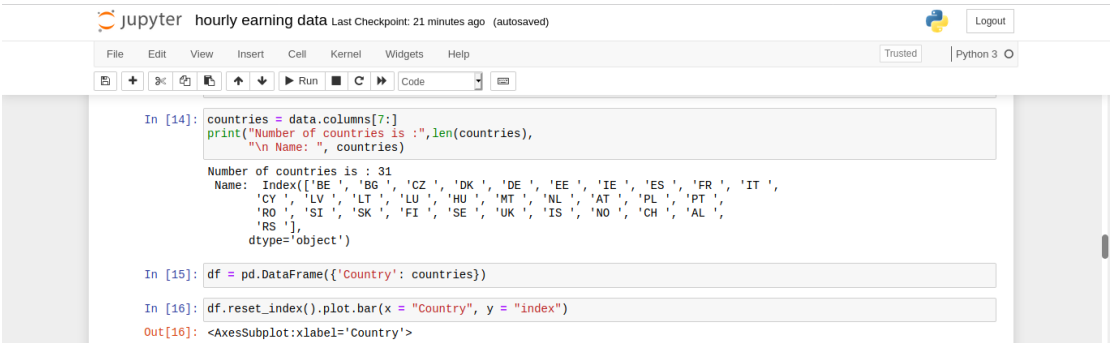


DATA ANALYSIS – VISUALIZATION

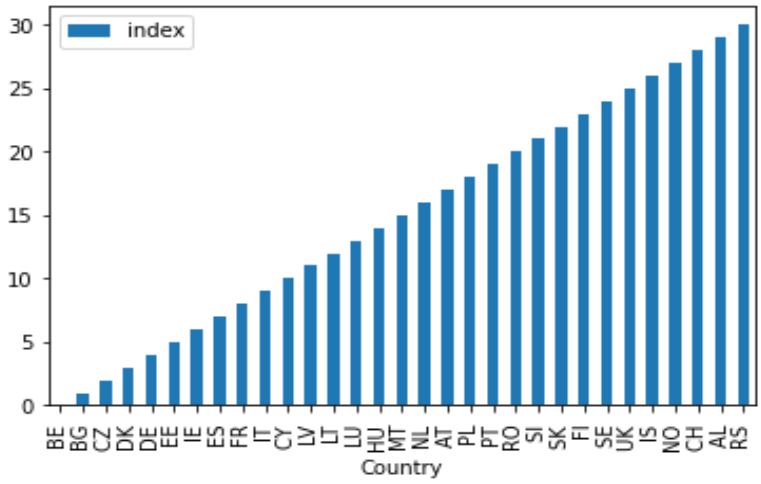
Code and its output with visualization

1. Which are the different countries from where data has been collected

Code:-

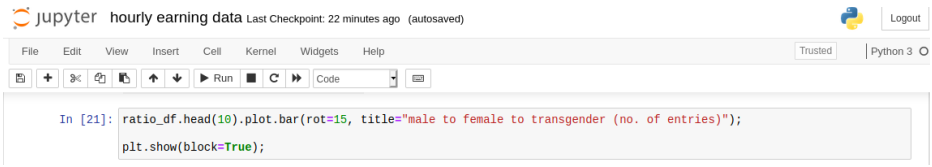


Output:-

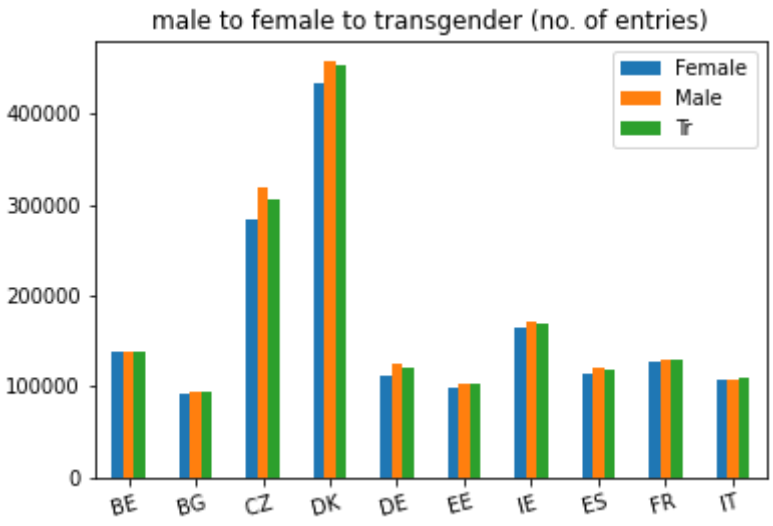


2. What is the male to female to transgender (no. of entries)

Code:-

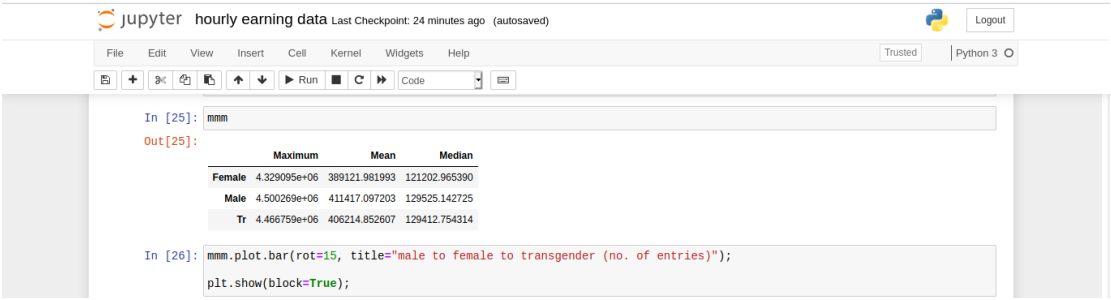


Output:-

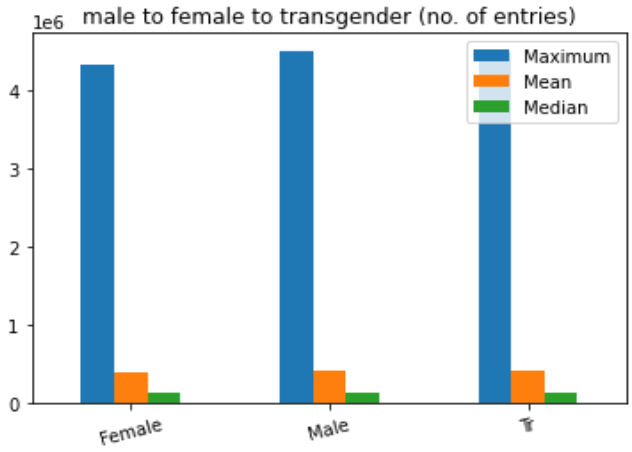


3. Max, mean and median income in 10 country for all different genders

Code:-

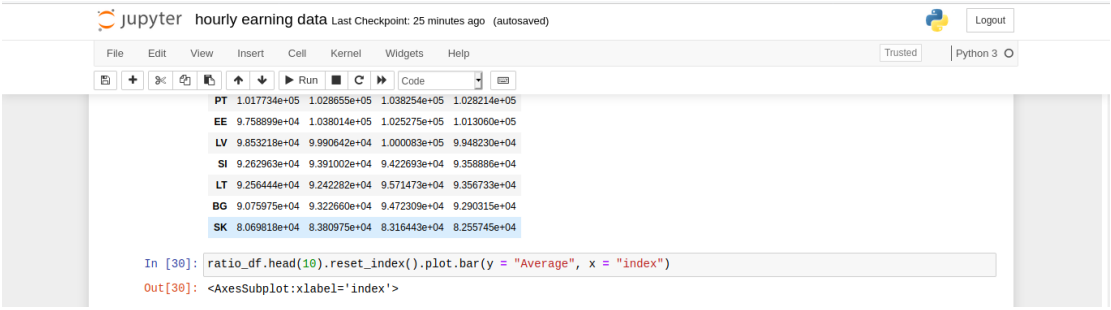


Output:-

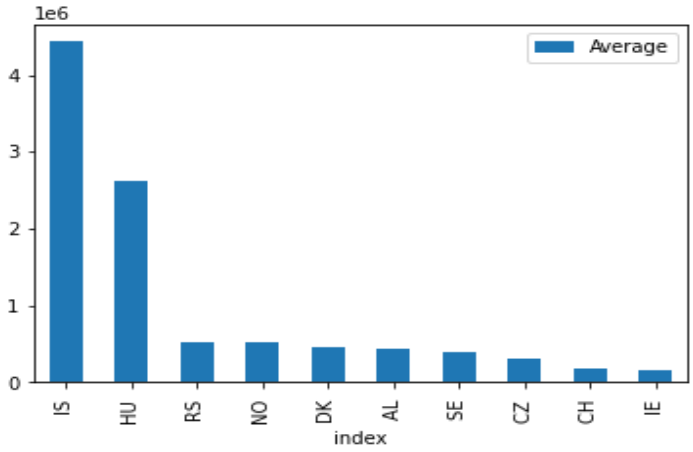


4. Average income in top 10 countries irrespective of the gender

Code:-



Output:-



5. What is the different types of sizeclas and their counts

Code:-

jupyter hourly earning data Last Checkpoint: 26 minutes ago (autosaved)

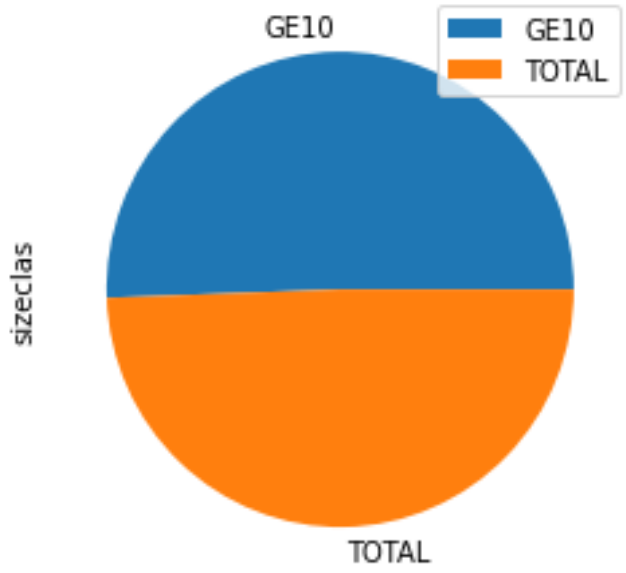
Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

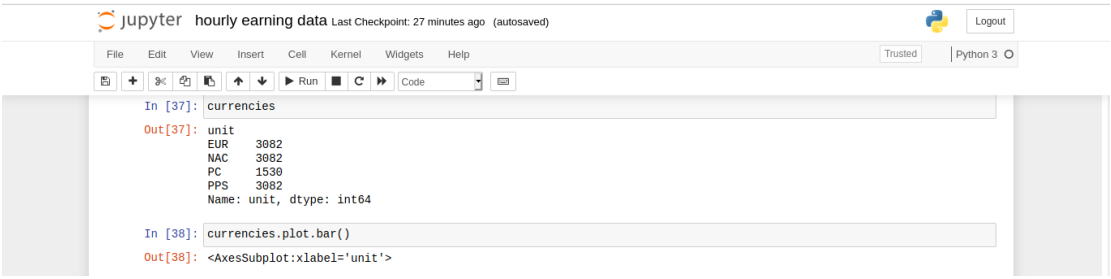
```
In [32]: unique_sizeclas = data.groupby('sizeclas')['sizeclas'].count()

In [33]: unique_sizeclas

Out[33]: sizeclas
         GE10     5449
         TOTAL    5327
         Name: sizeclas, dtype: int64

In [34]: unique_sizeclas.to_frame().plot.pie(y = 'sizeclas')

Out[34]: <AxesSubplot:ylabel='sizeclas'>
```

Output:-

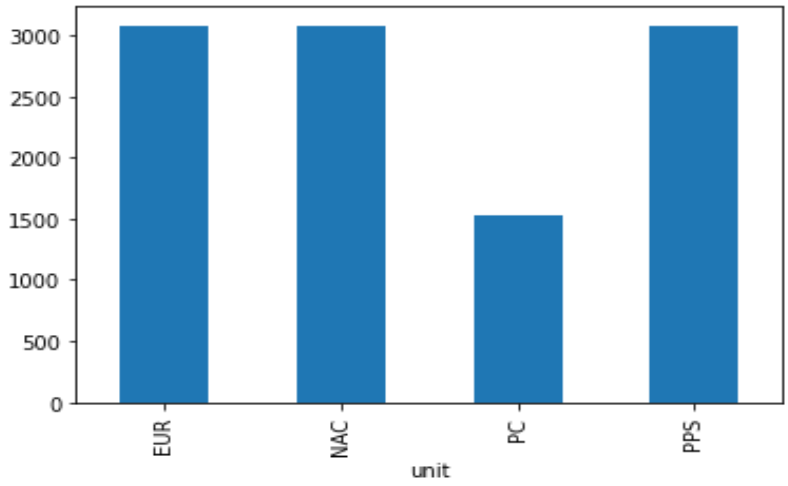


6. What is the different units of currencies used to pay and their counts

Code:-

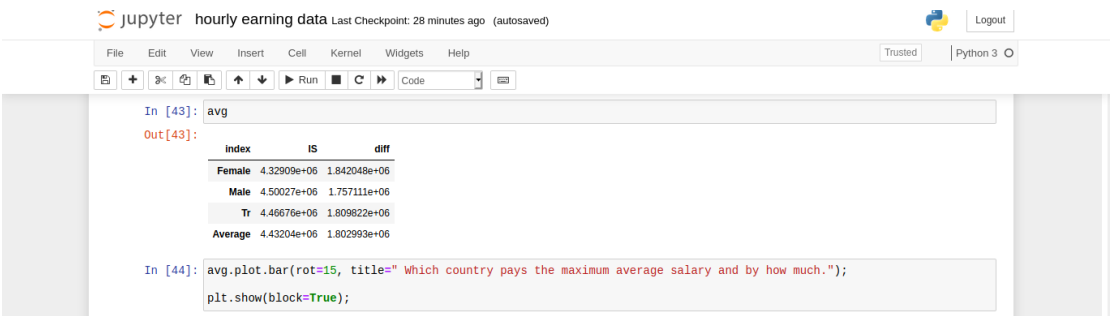


Output:-

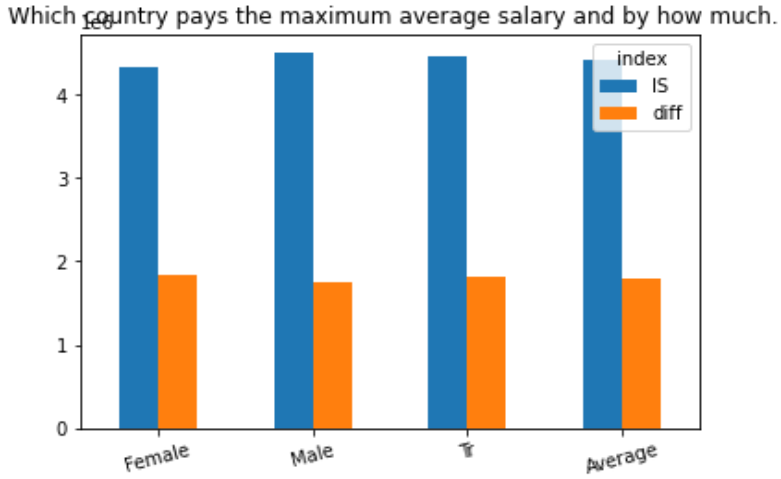


7. Which country pays the maximum average salary and by how much.

Code:-



Output:-



EXECUTIVE SUMMARY

CONCLUSION

In Europe, every country has similar hourly wages for each gender. Hourly wages of male is way more than female and transgenders. Also, Argentina has maximum hourly wages, and IS has the maximum average wage.

This data is based on the year 2018, so every conclusion we made is made wrt the year 2018.

Pros and cons..

PROS

- Helps completing small works with less pay.
- People can earn more as part time.
- Resolve financial issues.

CONS

- Hardworking.

REFERENCES

- https://www.thebalancecareers.com/hourly-vs-salary-employees-2063373
- https://data.europa.eu/euodp/en/data
- https://en.wikipedia.org/wiki/Wage