Author: Ganesh Kale

Week#1

Date: Dec 02, 2021

Assignment 1.2 Python Refresher

1. Import, Plot, Summarize, and Save Data

import required packages

In [1]:
```
import numpy as np  # for numeric operations
import pandas as pd  # for data manipulation
import matplotlib.pyplot as plt  # for data visualization
import seaborn as sns  # for data visualization


from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = 'all'
```

In [2]:
```
# load the dataset

occupation = pd.read_excel("Data/occupation.xlsx", sheet_name='Table 1.1',header=1)
```

In [3]:
```
# display head and shape of dataset

occupation.shape
occupation.head()
```

Out[3]:
```
(23, 7)
```

Out[3]:

	2020 National Employment Matrix title	2020 National Employment Matrix code	Employment, 2020	Employment, 2030	Employment change, 2020-30	Percent employment change, 2020-30	Median annual wage, 2020(1)
0	Total, all occupations	00-0000	153533.8	165413.7	11879.9	7.7	41950
1	Management occupations	11-0000	9782.3	10689.1	906.8	9.3	109760
2	Business and financial operations occupations	13-0000	9422.5	10173.3	750.8	8.0	72250
3	Computer and mathematical occupations	15-0000	5225.0	5959.9	734.9	14.1	91350
4	Architecture and engineering occupations	17-0000	2603.0	2748.9	146.0	5.6	83160

In [4]:
```
# drop first row since it is total

occupation.drop(index=occupation.index[0],inplace=True)
occupation.head(2)
```

Out[4]:

	2020 National Employment Matrix title	2020 National Employment Matrix code	Employment, 2020	Employment, 2030	Employment change, 2020-30	Percent employment change, 2020-30	Median annual wage, 2020(1)
1	Management occupations	11-0000	9782.3	10689.1	906.8	9.3	109760
2	Business and financial operations occupations	13-0000	9422.5	10173.3	750.8	8.0	72250

In [5]:
```
# display informaition about dataset

occupation.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 22 entries, 1 to 22
Data columns (total 7 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   2020 National Employment Matrix title  22 non-null     object
 1   2020 National Employment Matrix code   22 non-null     object
 2   Employment, 2020                       22 non-null     float64
 3   Employment, 2030                       22 non-null     float64
 4   Employment change, 2020-30             22 non-null     float64
 5   Percent employment change, 2020-30     22 non-null     float64
 6   Median annual wage, 2020(1)            22 non-null     int64
dtypes: float64(4), int64(1), object(2)
memory usage: 1.4+ KB
```

In [6]:
```
# display statistical summary of all numerical features from dataset

occupation.describe()
```

Out[6]:

	Employment, 2020	Employment, 2030	Employment change, 2020-30	Percent employment change, 2020-30	Median annual wage, 2020(1)
count	22.000000	22.000000	22.000000	22.000000	22.000000
mean	6978.813636	7518.813636	540.004545	9.068182	52732.272727
std	4756.275959	4849.918215	620.235147	6.697291	23861.960948
min	1061.800000	1088.400000	-539.200000	-2.800000	25500.000000
25%	2957.600000	3289.400000	123.950000	5.950000	32145.000000
50%	6342.550000	6805.500000	389.650000	8.600000	48065.000000
75%	9350.550000	10143.775000	890.350000	12.000000	69842.500000
max	19554.700000	19015.600000	2267.600000	23.100000	109760.000000

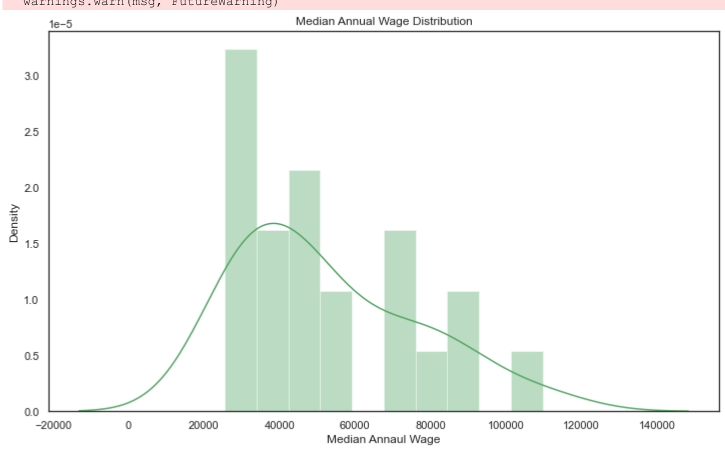
Visualization

In [7]:
```
# plot histogrma of annual wages to see distribution

sns.set(style='white')
plt.figure(figsize=(12,7))

sns.distplot(occupation['Median annual wage, 2020(1)'], color='g', bins=10)
plt.title('Median Annual Wage Distribution')
plt.xlabel('Median Annaul Wage')
plt.ylabel('Density');
```

```
/Users/ganeshkale/work/virtual_envs/venv/lib/python3.8/site-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

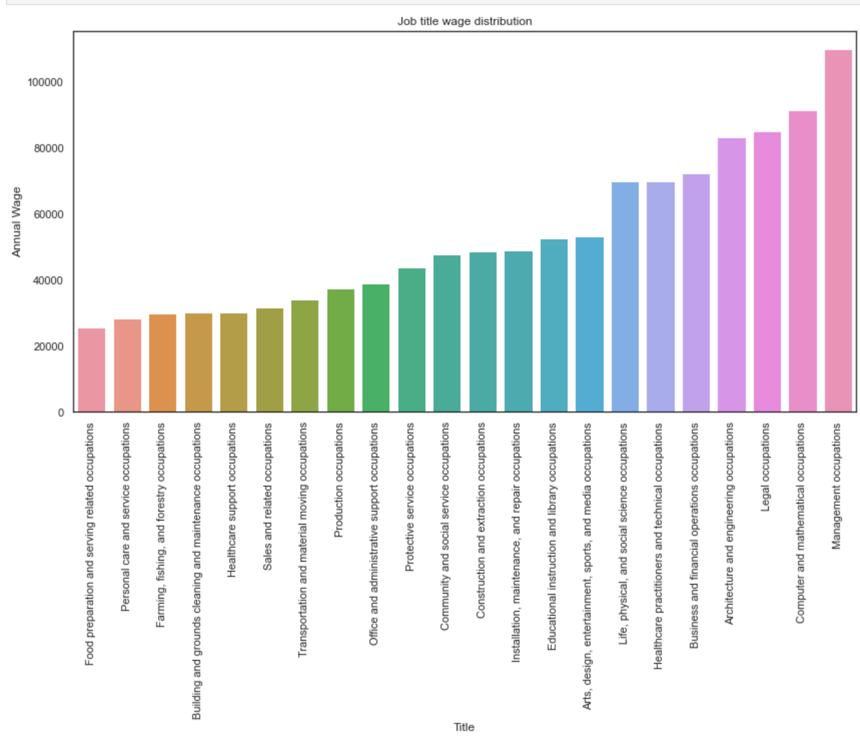


Based on the above histogram we see that the annual wages of the occupations with most job growth is not centrally distributed. The median annual salary for different occupations is around 40K. There are only one occupation with highest median annual wage above 100K and minimum wage is around 25K.

In [8]:
```
# plot bar plot to see annual wages by job title


sns.set(style='white')
plt.figure(figsize=(14,7))

sns.barplot(data=occupation.sort_values(by='Median annual wage, 2020(1)'),x='2020 National Employment Matrix ti
plt.title('Job title wage distribution')
plt.xlabel('Title')
plt.ylabel('Annual Wage')
plt.xticks(rotation=90);
```

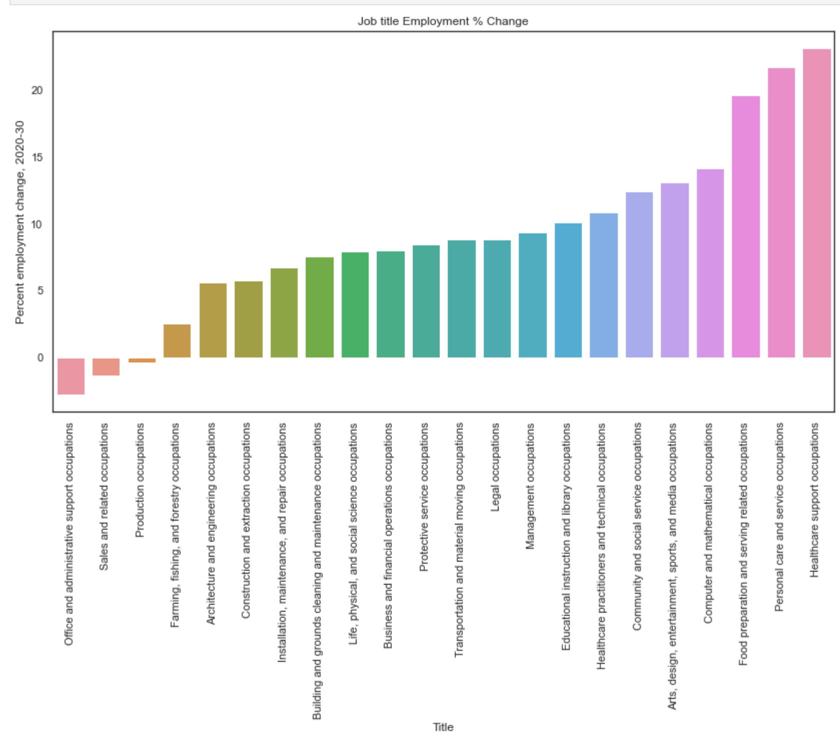


The above bar chart shows the distribution of different occupations and their annual wage, based on this we see that the management occupation has highest annual wage which is above 100K. The occupations such as social science, healthcare, financial operations,architecture and engineering,legal and computer professionals having annual wage range from 60K to 90K. There are 9 occupations with annual wage less thank 40K such as food prep, personal care, farming, grounds cleaning, healthcare support, sales, transportation, production etc. There are 6 occupations annual wage rages from 40K-60K such as social service, construction, maintenance, education and entertainment etc.

In [9]:
```
# plot bar plot to see annual wages by job title


sns.set(style='white')
plt.figure(figsize=(14,7))

sns.barplot(data=occupation.sort_values(by='Percent employment change, 2020-30'),x='2020 National Employment Ma
plt.title('Job title Employment % Change')
plt.xlabel('Title')
plt.ylabel('Percent employment change, 2020-30')
plt.xticks(rotation=90);
```



The above bar chart shows the percentage of emploment change in 2020 and 2030 by different job titles. Based on the bar chart we can say that the occupations such as office admins, sales and productions having negative change means the in 2030 the employment grwoth will go down. The occupations such as food prep, personal care, computer and healthe care support will have highest percentahe of change in employment.

Save Data to csv file

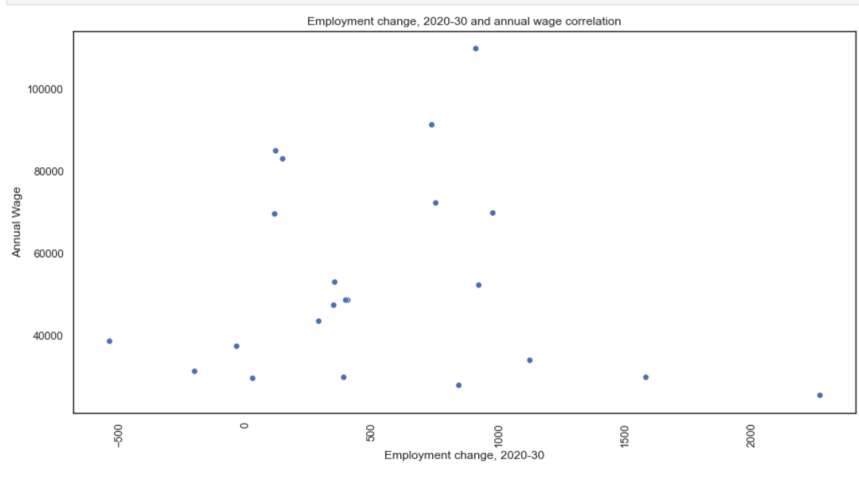
In [10]:
```
occupation.to_csv('occupations_with_Most_job_growth.csv', index=False)
```

2. Explore Some Bivariate Relations

In [11]:
```
# scatter plot to see relation between Employment, 2020 and Employment, 2030

sns.set(style='white')
plt.figure(figsize=(14,7))

sns.scatterplot(data=occupation,x='Employment change, 2020-30',y='Median annual wage, 2020(1)')
plt.title('Employment change, 2020-30 and annual wage correlation')
plt.xlabel('Employment change, 2020-30')
plt.ylabel('Annual Wage')
plt.xticks(rotation=90);
```

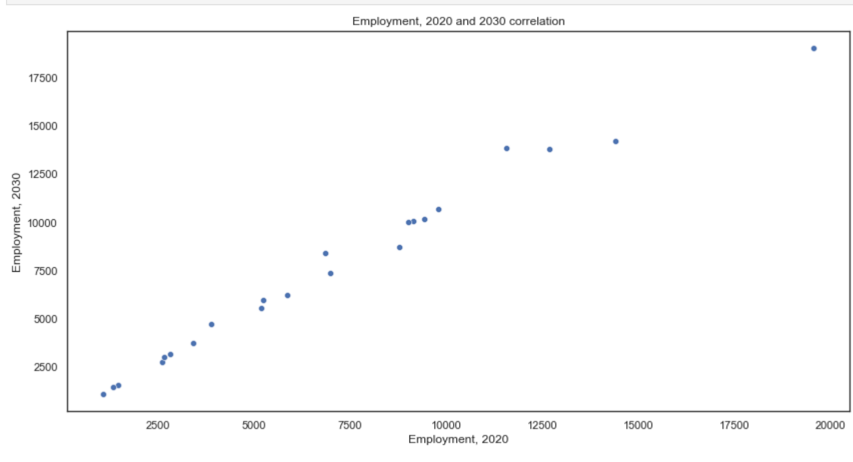


The above scatter plot of Employment change, 2020-30 vs Median Annual wage shows the correlation between them. Based on this chart we dont see the they dont have stromg correlation and both are independent to one naother.

In [12]:
```
# scatter plot to see relation between Employment, 2020 and Employment, 2030

sns.set(style='white')
plt.figure(figsize=(14,7))

sns.scatterplot(data=occupation,x='Employment, 2020',y='Employment, 2030')
plt.title('Employment, 2020 and 2030 correlation')
plt.xlabel('Employment, 2020')
plt.ylabel('Employment, 2030');
```



Based on above scatter plot we see that the projection made for employment 2030 is directly correlated with employment in 2020 and we see strong correaltion between them.

Correlation

In [13]:
```
occupation.corr()
```

Out[13]:

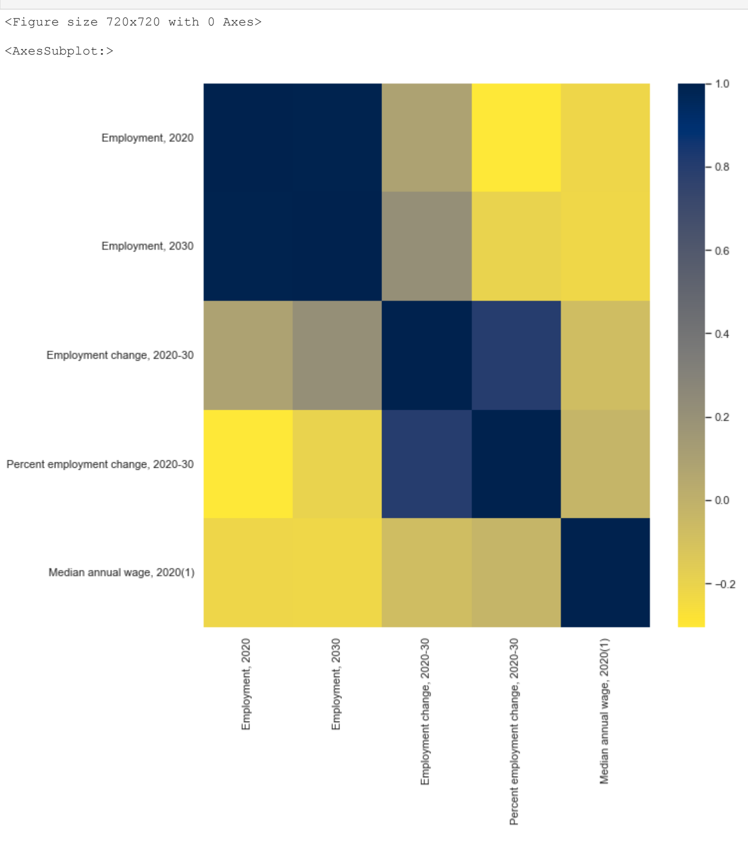
	Employment, 2020	Employment, 2030	Employment change, 2020-30	Percent employment change, 2020-30	Median annual wage, 2020(1)
Employment, 2020	1.000000	0.991852	0.087248	-0.306156	-0.215240
Employment, 2030	0.991852	1.000000	0.213447	-0.198126	-0.220345
Employment change, 2020-30	0.087248	0.213447	1.000000	0.798523	-0.072382
Percent employment change, 2020-30	-0.306156	-0.198126	0.798523	1.000000	-0.028619
Median annual wage, 2020(1)	-0.215240	-0.220345	-0.072382	-0.028619	1.000000

In [14]:
```
# plot heatmap of corr
sns.set(style='white')
plt.figure(figsize=(10,10))

sns.heatmap(occupation.corr(),cmap='cividis_r')
```

Out[14]:
```
<Figure size 720x720 with 0 Axes>
```

Out[14]:
```
<AxesSubplot:>
```



The baove correlation chart (heatmap) shows the correlation of different fields from the data set. The more darker the color is the more stroger correlation between the features, the diagonal shows the darker color since all features are directly correlated to each other.

We see employment change in 2020 is having stronger correaltion with employment change in 2030 since it is a projection from 2020 and it tells use that the projection is based on 2020.

We do not see any other features those are having stromg correaltion between them apart from employment change 2020 and 2030 so all of them are indepee

Organize a Data Report

Data Features Summary

Employment is projected to grow from 153.5 million to 165.4 million jobs from 2020 to 2030. Pandemic recovery and growth in healthcare-related occupations are expected to account for a large share of projected job growth. The data set shows the occupations with most job growth and tehir median annual wage.

Belwo are the feature details from this dataset:

- 2020 National Employment Matrix title: Title of the occupations those are expected to grow in 2030: Text Data Type
- 2020 National Employment Matrix code: The unique code of the occupations.: Categorical Data Type
- Employment, 2020: The number of jobs in 2020 in thousands: Numerical Data Type
- Employment, 2030: The number of jobs projected in 2030 in thousands: Numerical Data Type
- Employment change, 2020-30: The difference between field projected job count with current job count,i.e Employment, 2030 - Employment, 2020: Numerical Data Type
- Percent employment change, 2020-30: The percentage of employment change: Numerical Data Type
- Median annual wage, 2020(1): The median annual wage of the occupation in dollars: Currency/numerical Data Type

ack: Data are from the Occupational Employment and Wage Statistics program, U.S. Bureau of Labor Statistics. Wage data cover non-farm wage and salary workers and do not cover the self-employed, owners and partners in unincorporated firms, or household workers.

Conclusion:

The analysis performed on the dataset of Occupations with Most job growth by importing it from BLS website, analyzed the statitscal summary of the all the numerical variables from the dataset, checked the distribution of annual wage, analyzed the distibution of occupation with annual wage to see what occupation has highest median annaul wage and what occupation will have most percentage of change in emplyment in 2030. Analyzed the correlation of different variable and checked the correaltion coefficients.

END