PROJECT-1

EXPLORATORY ANALYSIS

DATASET: CITIES WITH BEST WORK-LIFE BALANCE

Cities with best work-life balance in 2022 Dataset : Cities with the Best Work-Life Balance 2022, This dataset consists of 24 attributes each of them contributing for the study of Work-life fit in various cities.The dataset consists of coloumns: 24 , rows : 100 , out of which 13 are of datatype float64, 2 are of int64 datatype and 9 are of object type data. 24 different coloumns are namely : 2022,Minimum Vacations Offered (Days),Unemployment,Covid Impact,Covid Support,Healthcare,Access to Mental Healthcare,Inclusivity&Tolerance,Affordability,Happiness-Culture & Leisure,City Safety,Outdoor Spaces,Air Quality,Wellness and Fitness,TOTAL SCORE,Paid Parental Leave (Days),Inflation,Multiple

Jobholders,Vacations Taken (Days),Overworked Population,2021,City,Remote Jobs,Country etc... Cities with the Best Work-Life Balance 2022 1.Three aspects of work-life balance 2.Work Intensity 3.Society and Institutions 4.City Liveability.

Work-life balance entails to a balanced state, where one adequately balances work or professional demands and those of their personal life. An individual who lacks a work-life balance has more obligations with respect to work and home, works longer hours, and experiences shortfall in personal time.Some utilize work-life balance as an opportunity to work no more than 8 hours a day and still have time to hit the gym, run some other works, and spend time with family and friends.Here we get to se which cities best aid work-life balance.

Not maintaining proper work-life blend can lead to mental health issues such as depression, anxiety, and insomnia, as well as physical health issues including chronic aches and pains, heart troubles, and hypertension. Burnout happens when an employee suffers too much stress over a long period of time.

Data preparation :

Replacing the irrelevant data in the dataset.

```
In [7]: df['Covid Support']=df['Covid Support'].replace('-','0')

In [8]: df['Remote Jobs']=df['Remote Jobs'].replace('%','',regex=True).astype(float)

In [9]: df['Overworked Population']=df['Overworked Population'].replace('%','',regex=True).astype(float)

In [10]: df['Minimum Vacations Offered (Days)']=df['Minimum Vacations Offered (Days)'].replace('%','',regex=True).astype(float)

In [11]: df['Vacations Taken (Days)']=df['Vacations Taken (Days)'].replace('-','0')
         df['Vacations Taken (Days)']=df['Vacations Taken (Days)'].replace('%','',regex=True).astype(float)

In [12]: df['Inflation']=df['Inflation'].replace('%','',regex=True).astype(float)

In [13]: df['Multiple Jobholders']=df['Multiple Jobholders'].replace('%','',regex=True).astype(float)
```

Removing the null values from the dataset

```
In [55]: df1 = df.mask(df == '0.0', np.nan)

In [17]: df1 = df.dropna()
```

However the prevailant null values and the irrelevant data is hardly about 2-5% of the dataset as a whole which doesnot effect the further analysis.

Getting the information of the attributes and their statstical report respectively

```
In [52]: df1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 24 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   2022                               100 non-null    int64
 1   2021                               100 non-null    object
 2   City                               100 non-null    object
 3   Country                            100 non-null    object
 4   Remote Jobs                        100 non-null    float64
 5   Overworked Population              100 non-null    float64
 6   Minimum Vacations Offered (Days)   100 non-null    float64
 7   Vacations Taken (Days)             100 non-null    float64
 8   Unemployment                       100 non-null    float64
 9   Multiple Jobholders                100 non-null    float64
 10  Inflation                          100 non-null    float64
 11  Paid Parental Leave (Days)         100 non-null    object
 12  Covid Impact                       100 non-null    float64
 13  Covid Support                      100 non-null    float64
 14  Healthcare                         100 non-null    float64
 15  Access to Mental Healthcare        100 non-null    float64
 16  Inclusivity & Tolerance            100 non-null    float64
 17  Affordability                      100 non-null    float64
 18  Happiness, Culture & Leisure       100 non-null    float64
 19  City Safety                        100 non-null    float64
 20  Outdoor Spaces                     100 non-null    float64
 21  Air Quality                        100 non-null    float64
 22  Wellness and Fitness               100 non-null    float64
 23  TOTAL SCORE                        100 non-null    float64
dtypes: float64(19), int64(1), object(4)
memory usage: 18.9+ KB
```

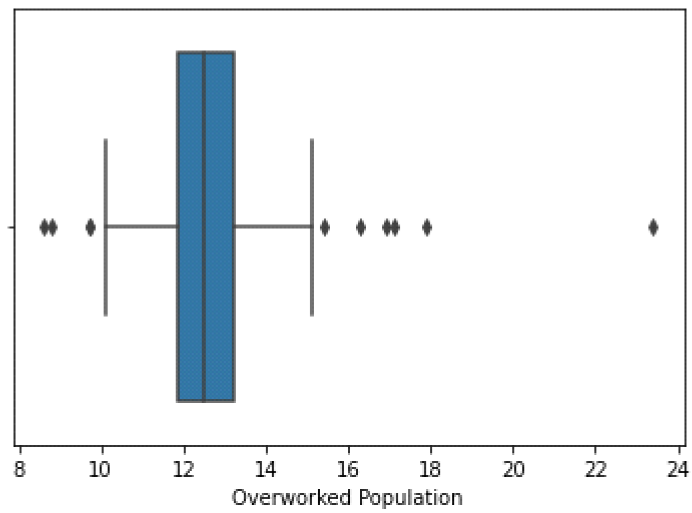
```
In [15]: df.describe()
Out[15]:
```

	2022	Remote Jobs	Overworked Population	Minimum Vacations Offered (Days)	Vacations Taken (Days)	Unemployment	Multiple Jobholders	Inflation	Covid Impact	Covid Support	Healthcare	Access to Mental Healthcare	In T
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.00000	100.000000	100.000000	
mean	50.500000	37.841600	12.645000	14.410000	14.647000	93.292000	5.016000	8.548400	77.875000	84.95300	88.753000	72.143000	
std	29.011492	5.083966	1.866146	6.313518	8.044641	5.888556	1.758944	10.310558	6.959949	6.91823	8.483537	8.688286	
min	1.000000	16.840000	8.600000	6.000000	0.000000	50.000000	1.100000	0.890000	50.000000	50.00000	50.000000	50.000000	
25%	25.750000	36.237500	11.875000	10.000000	8.700000	92.675000	3.800000	6.260000	74.150000	81.10000	86.000000	66.600000	
50%	50.500000	37.775000	12.500000	10.000000	9.400000	94.850000	4.800000	7.735000	78.350000	84.90000	89.000000	67.500000	
75%	75.250000	41.180000	13.200000	20.000000	24.025000	95.950000	6.000000	9.050000	82.350000	89.45000	94.800000	78.600000	
max	100.000000	52.060000	23.400000	30.000000	30.000000	100.000000	10.000000	107.410000	100.000000	100.00000	100.000000	100.000000	

Doing the outlier analysis

Outlier analysis is performed by plotting a boxplot.Inspite of having the outliers it does not make significant changes to the analysis. As the attributes in the dataset we chose have varied behaviour i.e. they are not set to behave a certain way , removing the outliers is not considered in this case.

```
In [45]: #plotting a boxplot for outliers
         sns.boxplot(df['Overworked Population'])

Out[45]: <AxesSubplot:xlabel='Overworked Population'>
```

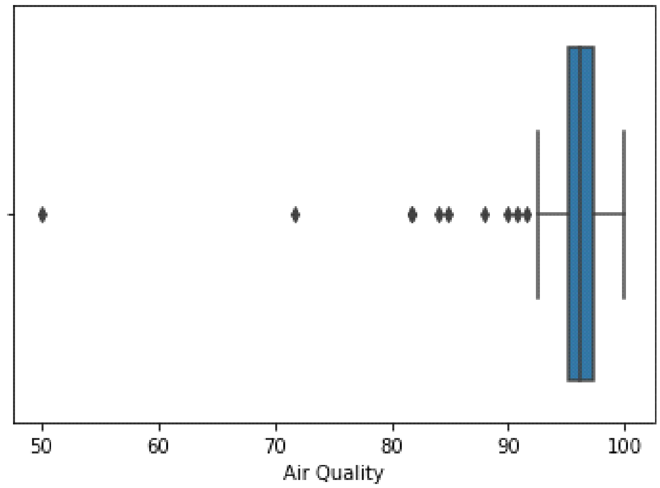



```
In [46]: #identifying the outliers
         np.where(df['Overworked Population']<10)

Out[46]: (array([ 7, 11, 12, 94], dtype=int64),)
```

```
In [47]: #identifying the outliers
         np.where(df['Overworked Population']>15)

Out[47]: (array([13, 44, 92, 93, 95, 97, 98], dtype=int64),)
```

In [36]:
```
#plotting a boxplot for outliers
sns.boxplot(df['Air Quality'])
```

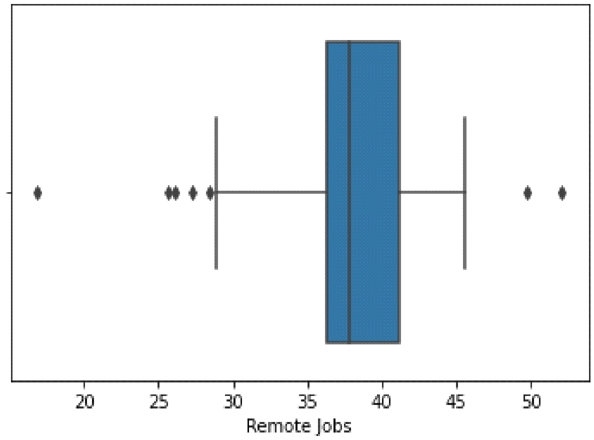
Out[36]: <AxesSubplot:xlabel='Air Quality'>



In [37]:
```
#identifying the outliers
np.where(df['Air Quality']<91)
```

Out[37]: (array([13, 44, 87, 90, 92, 95, 96, 97, 98], dtype=int64),)

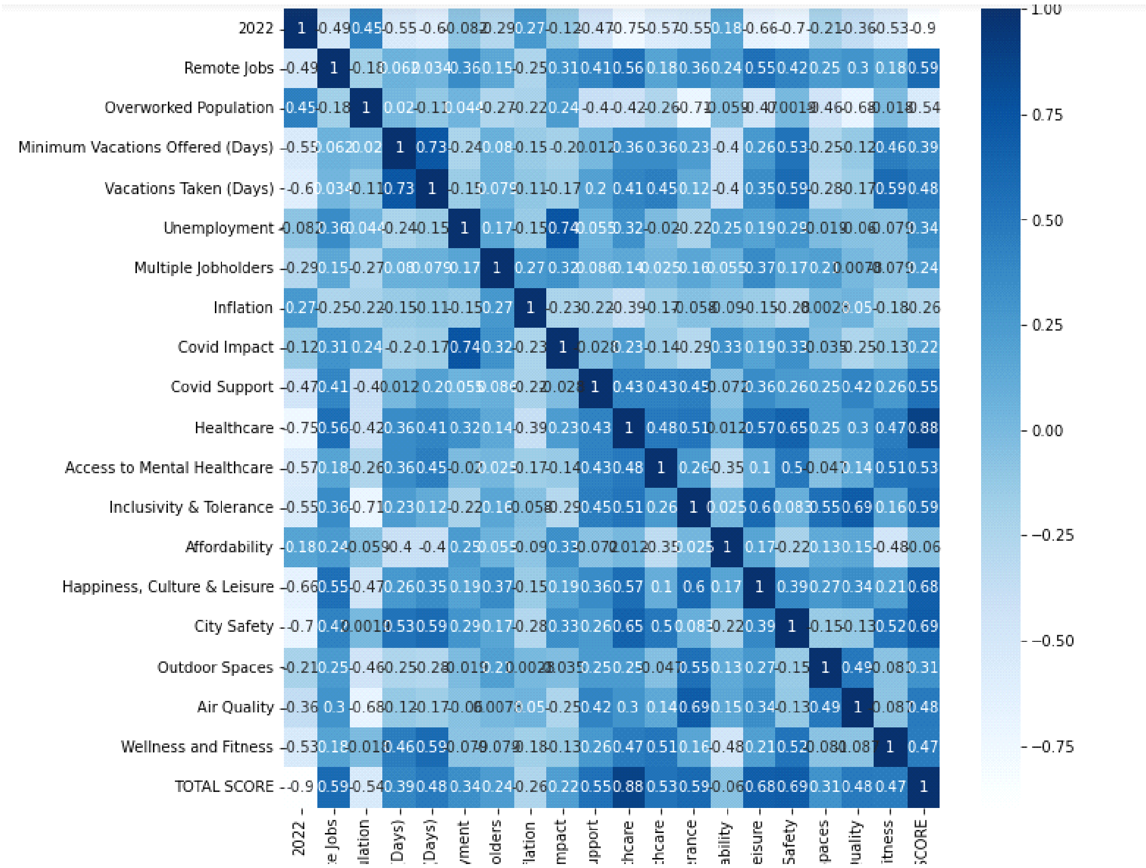
```
In [42]: #plotting a boxplot for outliers
         sns.boxplot(df['Remote Jobs'])

Out[42]: <AxesSubplot:xlabel='Remote Jobs'>
```




```
In [43]: np.where(df['Remote Jobs']<28)

         #identifying the outliers

Out[43]: (array([93, 95, 96, 99], dtype=int64),)

In [44]: #identifying the outliers
         np.where(df['Remote Jobs']>45)

Out[44]: (array([40, 44, 61], dtype=int64),)
```

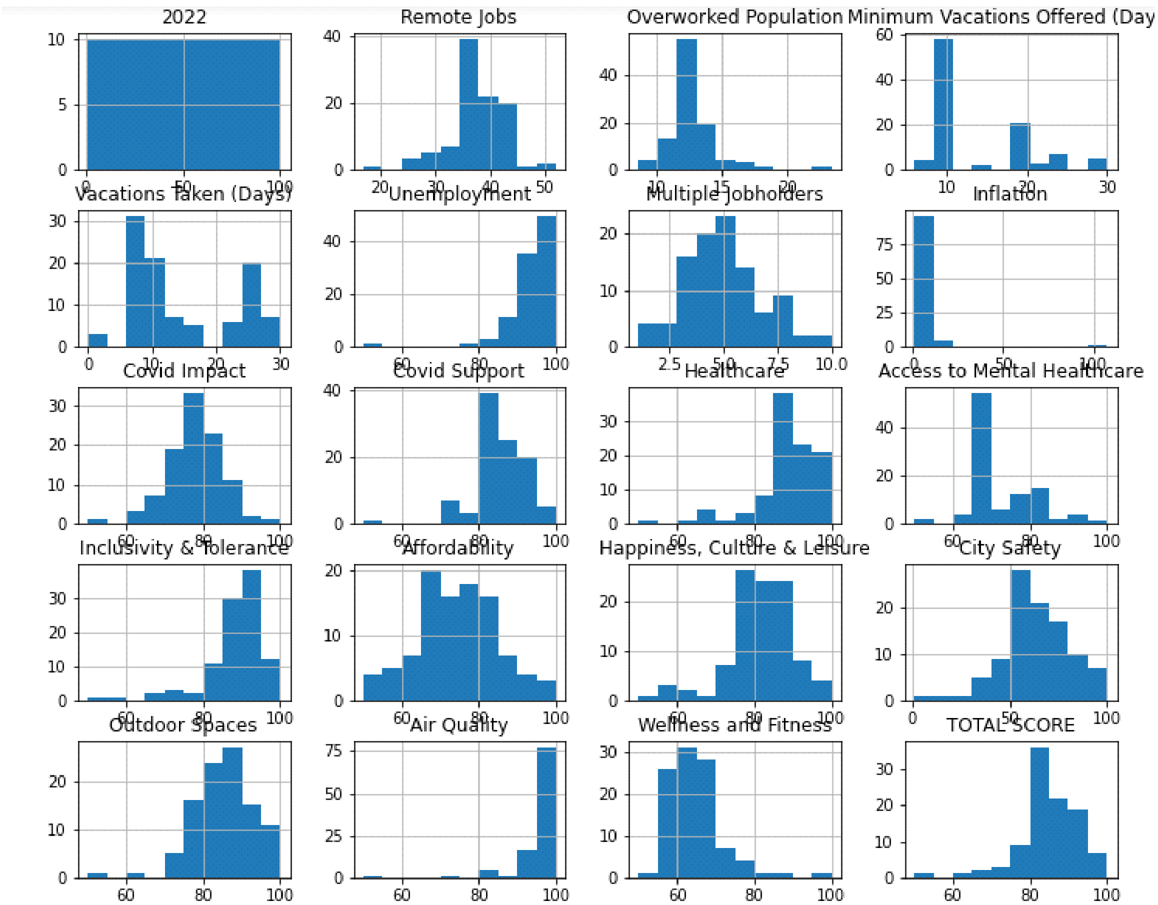
Exploratory analysis

*Plotting a heatmap showing the correlation between the attributes .



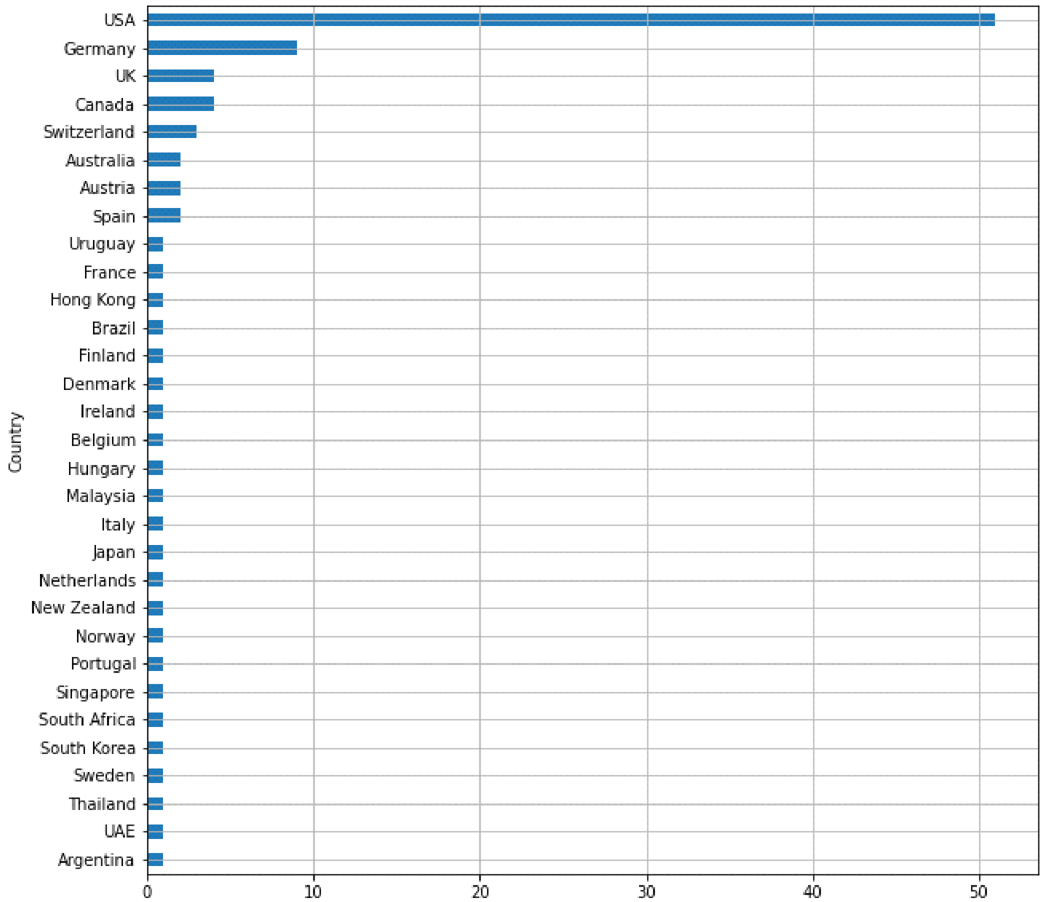
We see high correlation between Healthcare and TotalSCORE, and Unempoyment with Covid Impact

*Plotting and visualizing the data in the form of a histogram.

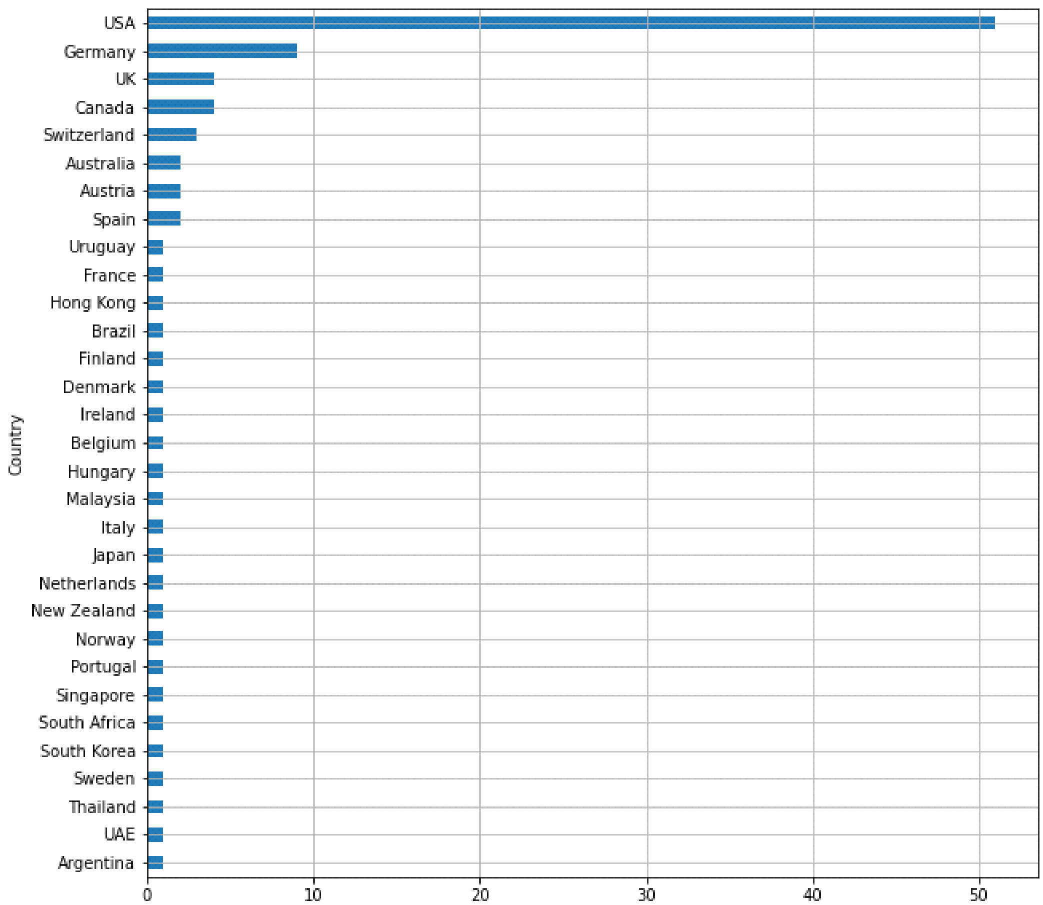


By performing analysis between various attributes the following inferences are made.

* Most number of cities for the survey are considered from the U.S , Germany , and Canada respectively according to the graph below.

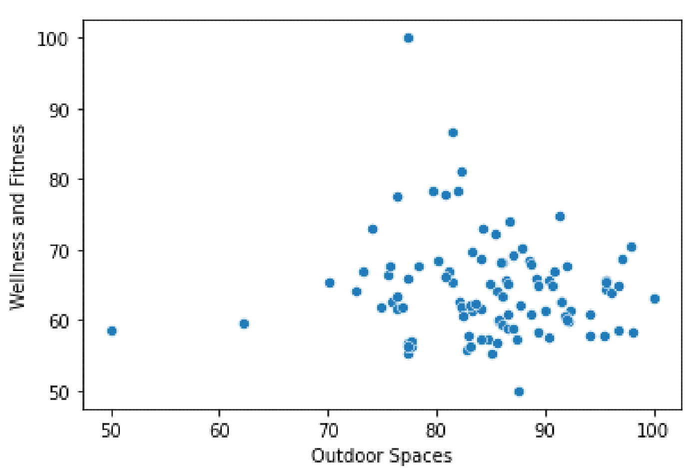


*It is also seen that U.S has most of the remote jobs then Germany UK,Canada and so on.



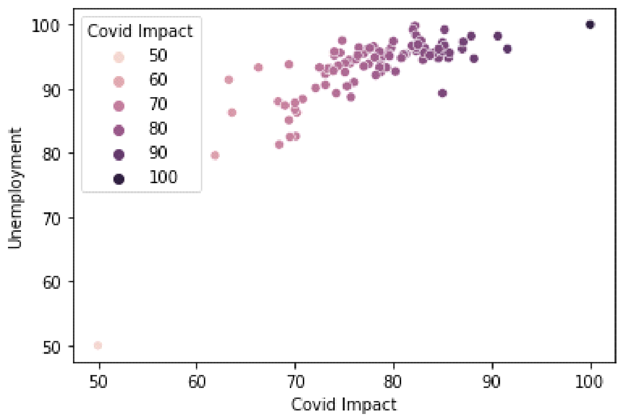
*Wellness and fitness of people are not highly affected by the availability of outdoor spaces.

Out[13]: <AxesSubplot:xlabel='Outdoor Spaces', ylabel='Wellness and Fitness'>

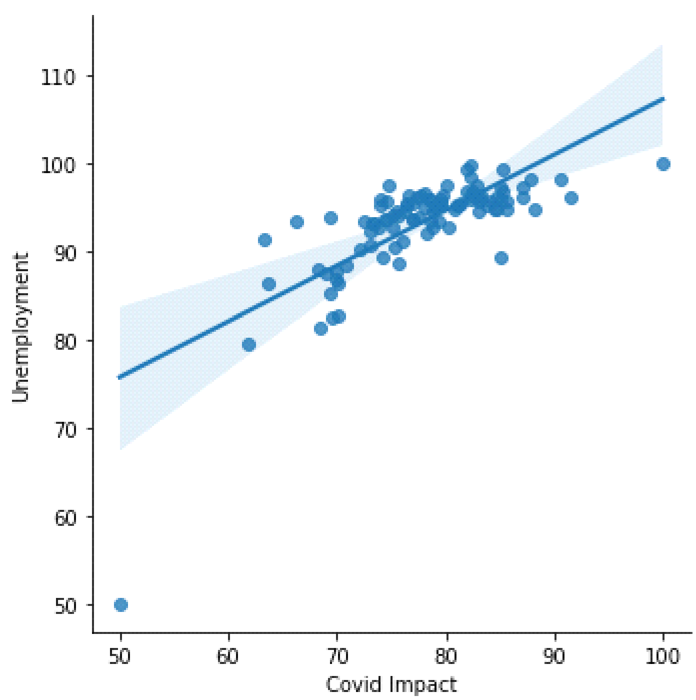


*It is evident that covid impact has great influence on unemployement,there is a direct impact of covid on employment .

Out[5]: <AxesSubplot:xlabel='Covid Impact', ylabel='Unemployment'>

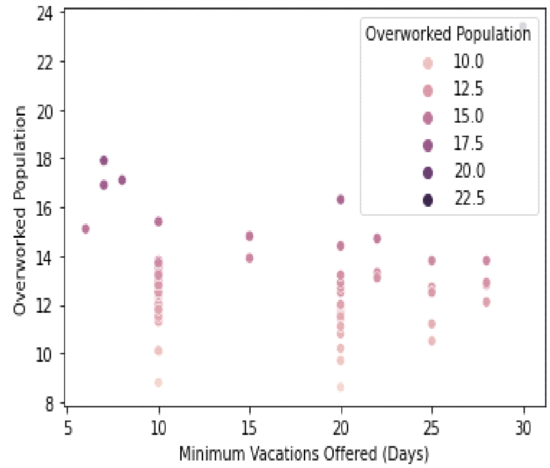


Out[50]: <seaborn.axisgrid.FacetGrid at 0x1e3e8325d90>



*Minimum Vacations Offered (Days) are hardly more than 20 days for the population who are considered overworked.

Out[48]: <AxesSubplot:xlabel='Minimum Vacations Offered (Days)', ylabel='Overworked Population'>



*As the percentage of Overworked population increases the count in the Happiness, Culture & Leisure subsequently decrease.

Out[51]: <AxesSubplot:xlabel='Overworked Population', ylabel='Happiness, Culture & Leisure'>