Machine Learning Course

Finall project

Yasna Valipour

Fall 2023

معرفی دیتاست

دیتاست خام شامل اطلاعات مربوط به 1000 دانشگاه برتر جهان میباشد. نمای کلی دیتا به صورت زیر میباشد که شامل 12 ستون و 1000 سطر میباشد.

	World Rank	Institution	Location	National Rank	Quality of Education	Alumni Employment	Quality of Faculty	Research Output	Quality Publications	Influence	Citations	Score
0	1	Harvard University	USA	1	2	1	1	1	1	1	1	100.0
1	2	Stanford University	USA	2	10	3	2	10	4	3	2	96.7
2	3	Massachusetts Institute of Technology	USA	3	3	11	3	30	15	2	6	95.1
3	4	University of Cambridge	United Kingdom	1	5	19	6	12	8	6	19	94.0
4	5	University of Oxford	United Kingdom	2	9	25	10	9	5	7	4	93.2
...	...	...	...	...	...	...	...	...	...	...	...	...
995	996	Aga Khan University	Pakistan	3	-	> 1000	-	> 1000	> 1000	464	673	69.8
996	997	University of Calcutta	India	17	353	716	296	798	966	> 1000	> 1000	69.8
997	998	K?chi University	Japan	56	-	> 1000	-	> 1000	> 1000	811	673	69.8
998	999	Soonchunhyang University	South Korea	35	-	> 1000	-	881	> 1000	> 1000	898	69.8
999	1000	Capital Normal University	China	108	-	869	-	923	904	889	> 1000	69.8

1000 rows × 12 columns

معرفی ستون ها

World Rank	Institution	Location	National Rank	Quality of Education	Alumni Employment	Quality of Faculty	Research Output	Quality Puplication	Influeance	Citations	Score
رتبه جهانی	موسسه	مکان جغرافیایی	رتبه ملی	کیفیت آموزش	استخدام فارغ التحصیلان	کیفیت دانشکده	خروجی تحقیق	انتشارات با کیفیت	نفوذ	استناد	امتیاز

```
# Data profilling
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   World Rank           1000 non-null   int64  
 1   Institution          1000 non-null   object 
 2   Location             1000 non-null   object 
 3   National Rank        1000 non-null   int64  
 4   Quality of Education  403 non-null    object 
 5   Alumni Employment    1000 non-null   object 
 6   Quality of Faculty    269 non-null    object 
 7   Research Output      1000 non-null   object 
 8   Quality Publications  1000 non-null   object 
 9   Influence            1000 non-null   object 
 10  Citations            1000 non-null   object 
 11  Score                1000 non-null   float64
dtypes: float64(1), int64(2), object(9)
```

فرآیند آماده سازی دیتا

همانطور که مشخص است در دو ستون کیفیت آموزش و کیفیت دانشکده بعضی از داده ها خالی هستند و چون مقدار آن ها زیاد است و نمیتوان با مقادیر احتمالی پر کرد بهتر است آن ها را NaN قرا دهیم. و نیز مقدار <1000 نیز دیده میشود که برای پردازش باید آنها را به مقدار عددی تبدیل کرد که در اینجا به فرض همه آنهارا 1001 قرار میدهیم.

```
# Replace values equal to'-' with NAN
df1 = df.replace('-', np.nan)

# Replace missing value
df2 = df1.replace('> 1000', 1001)
```

با انجام این دستوراتدیتاست ما به صورت زیر در می آید.

	World Rank	Institution	Location	National Rank	Quality of Education	Alumni Employment	Quality of Faculty	Research Output	Quality Publications	Influence	Citations	Score
0	1	Harvard University	USA	1	2	1	1	1	1	1	1	100.0
1	2	Stanford University	USA	2	10	3	2	10	4	3	2	96.7
2	3	Massachusetts Institute of Technology	USA	3	3	11	3	30	15	2	6	95.1
3	4	University of Cambridge	United Kingdom	1	5	19	6	12	8	6	19	94.0
4	5	University of Oxford	United Kingdom	2	9	25	10	9	5	7	4	93.2
...	...	...	...	...	...	...	...	...	...	...	...	...
995	996	Aga Khan University	Pakistan	3	NaN	1001	NaN	1001	1001	464	673	69.8
996	997	University of Calcutta	India	17	353	716	296	798	966	1001	1001	69.8
997	998	K?chi University	Japan	56	NaN	1001	NaN	1001	1001	811	673	69.8
998	999	Soonchunhyang University	South Korea	35	NaN	1001	NaN	881	1001	1001	898	69.8
999	1000	Capital Normal University	China	108	NaN	869	NaN	923	904	889	1001	69.8

1000 rows × 12 columns

اکنون اسم بعضی از ستون هارا به دلیل داشتن کارکتر مخفی تغییر میدهیم تا در پردازش مشکلی پیش نیاید.

```
# Rename columns to remove hidden characters

df2.rename(columns={'World Rank': 'World_Rank'}, inplace=True)
df2.rename(columns={'National Rank': 'National_Rank'}, inplace=True)
df2.rename(columns={'Alumni Employment': 'Alumni_Employment'}, inplace=True)
df2.rename(columns={'Research Output': 'Research_Output'}, inplace=True)
df2.rename(columns={'Quality Publications': 'Quality_Publications'}, inplace=True)
df2.columns.values[4] = 'Quality_of_Education'
df2.columns.values[6] = 'Quality_of_Faculty'
```

اکنون مقادیر میسینگ ولیو هارا به دست می آوریم.

```
# Missing Values

# Verifing
df.isnull().sum()
```

```
World_Rank                0
Institution               0
Location                  0
National_Rank             0
Quality_of_Education    597
Alumni_Employment         0
Quality_of_Faculty      731
Research_Output           0
Quality_Publications      0
Influence                 0
Citations                 0
Score                     0
dtype: int64
```

همان طور که مشاهده میکنیم در دو ستون مقادیر میسینگ ولیووها زیاد بوده و هیچ اطلاعاتی در دست نداریم که آنهارا پر کنیم و حذف سطر آنها دیتای مارا خراب میکند پس راه بهینه این است که این دو ستون را کلا حذف کرد.

```
# Delete two columns Quality of Education and Quality of Faculty

df3 = df2.drop(columns = ["Quality_of_Education"])
df4 = df3.drop(columns = ["Quality_of_Faculty"])
```

```
df4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   World_Rank            1000 non-null   int64  
 1   Institution           1000 non-null   object 
 2   Location              1000 non-null   object 
 3   National_Rank         1000 non-null   int64  
 4   Alumni_Employment     1000 non-null   object 
 5   Research_Output       1000 non-null   object 
 6   Quality_Publications  1000 non-null   object 
 7   Influence             1000 non-null   object 
 8   Citations             1000 non-null   object 
 9   Score                 1000 non-null   float64
dtypes: float64(1), int64(2), object(7)
```

طبق تصویر بالا برای آنالیز نموداری باید مقادیر آبجکت را در ستون های Quality_Publications , Influence , Citations , Research_Output , Alumni_Employment به اینتیجر یا فلوت تغییر دهیم.

```
# convert data

selected_columns = ['Alumni_Employment'
                    , 'Research_Output', 'Quality_Publications','Influence','Citations']
df4[selected_columns] = df4[selected_columns].apply(pd.to_numeric)
```

```
df4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   World_Rank            1000 non-null   int64
 1   Institution           1000 non-null   object
 2   Location              1000 non-null   object
 3   National_Rank         1000 non-null   int64
 4   Alumni_Employment     1000 non-null   int64
 5   Research_Output       1000 non-null   int64
 6   Quality_Publications  1000 non-null   int64
 7   Influence             1000 non-null   int64
 8   Citations             1000 non-null   int64
 9   Score                 1000 non-null   float64
dtypes: float64(1), int64(7), object(2)
```

برای آنالیز بهتر دیتاستمان یک ستون به عنوان قاره به دیتاست اضافه میکنیم.

دو ستون دیگر نیز با نام طول جغرافیایی و عرض جغرافیایی نیز اضافه میکنیم.

```
# add new column for continent

def get_continent_name(country_name):
    country_code = pc.country_name_to_country_alpha2(country_name, cn_name_format="default")
    continent_code = pc.country_alpha2_to_continent_code(country_code)
    return pc.convert_continent_code_to_continent_name(continent_code)

df4['Continent'] = df4['Location'].apply(get_continent_name)
```

```
# Add latitude and longitude columns

geolocator = Nominatim(user_agent="Yasna")

df4['Location'] = df4['Location'].apply(geolocator.geocode)  # Get geographic coordinates
df4['Latitude'] = df4['Location'].apply(lambda loc: loc.latitude if loc else None)  # Latitude extraction
df4['Longitude'] = df4['Location'].apply(lambda loc: loc.longitude if loc else None)  # Longitude extraction
```

سپس در نهایت برسی میکنیم که میسینگ ولیوییدر دیتاست ما وجود نداشته باشد.

```
df4.isna().sum()
```

```
World_Rank              0
Institution             0
Location                0
National_Rank           0
Alumni_Employment       0
Research_Output         0
Quality_Publications    0
Influence               0
Citations               0
Score                   0
Continent               0
Latitude                0
Longitude               0
dtype: int64
```

در نهایت

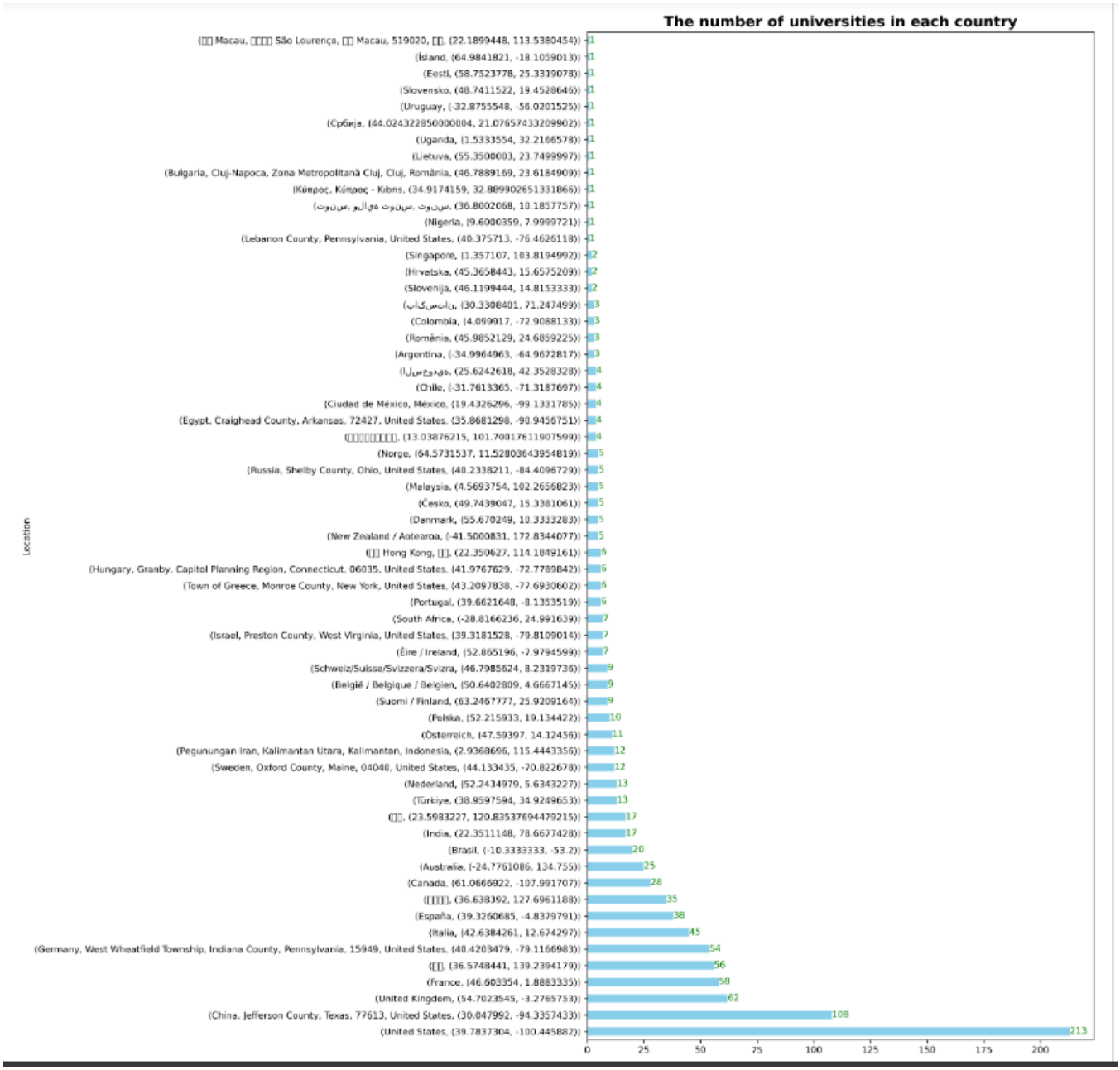
```
df4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   World_Rank            1000 non-null   int64  
 1   Institution           1000 non-null   object 
 2   Location              1000 non-null   object 
 3   National_Rank         1000 non-null   int64  
 4   Alumni_Employment     1000 non-null   int64  
 5   Research_Output       1000 non-null   int64  
 6   Quality_Publications  1000 non-null   int64  
 7   Influence             1000 non-null   int64  
 8   Citations             1000 non-null   int64  
 9   Score                 1000 non-null   float64
 10  Continent             1000 non-null   object 
 11  Latitude              1000 non-null   float64
 12  Longitude             1000 non-null   float64
dtypes: float64(3), int64(7), object(3)
```

اکنون دیتاست آماده پردازش میباشد.

پردازش دیتاست

_ تعداد دانشگاه ها در هر کشور



با توجه به این نمودار در بین 1000 دانشگاه برتر آمریکا با داشتن 213 دانشگاه در رتبه اول جهان قرار دارد و بعد آن چین با 108 دانشگاه و بعد انگلیس با 62 دانشگاه در رتبه های دوم و سوم قرار دارند.

کشور ما ایران نیز با داشتن 12 دانشگاه در رتبه پانزدهم قرار دارد.

ـ تاثیر عوامل مختلف بر روی امتیاز دانشگاه ها

World_Rank vs Score

National_Rank vs Score

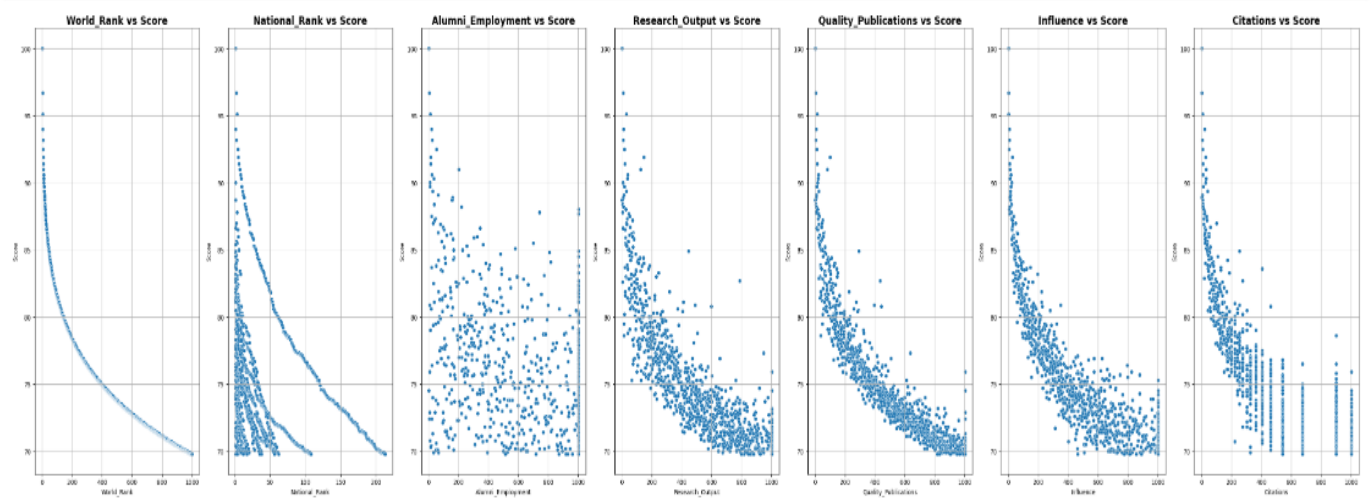
Alumni_Employment vs Score

Research_Output vs Score

Quality_Publications vs Score

Influence vs Score

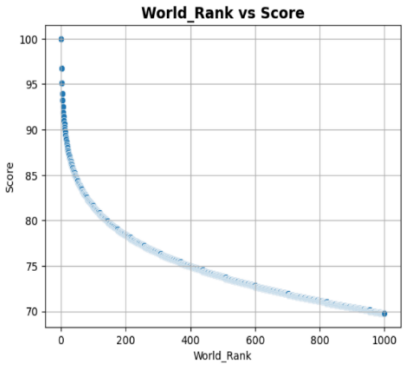
Citations vs Score



حال یکی یکی به برسی نمودار ها میپردازیم.

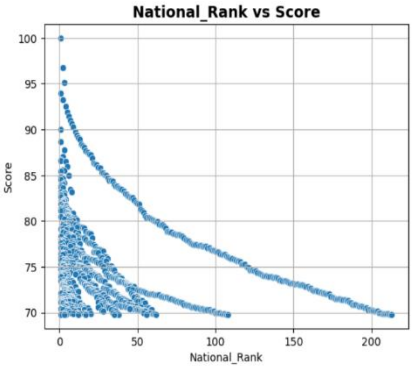
World_Rank vs Score-1

طبق نمودار پراکندگی روبرو هر چه امتیاز دانشگاه بیشتر باشد رتبه جهانی کمتری دارد.

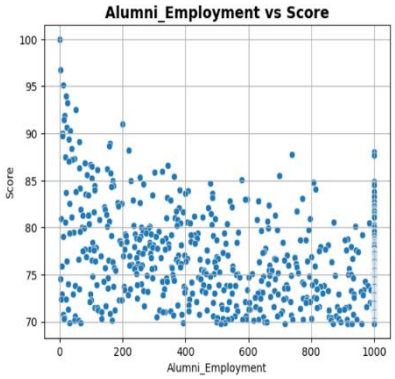


National_Rank vs Score-2

طبق نمودار روبرو نیز همانند رتبه جهانی، رتبه ملی نیز با امتیاز رابطه عکس دارد.

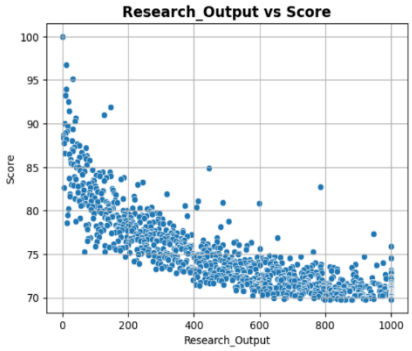


Alumni_Employment vs Score-3



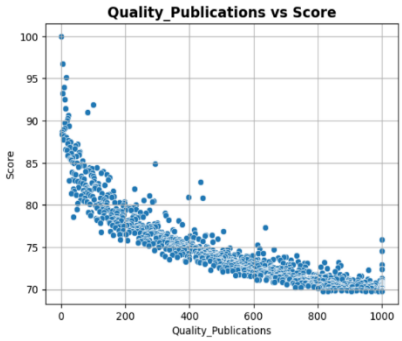
ویژگی استخدام فارغ التحصیلان دارای پراکندگی و همبستگی زیاد با سایر ویژگی هاست و حاوی امتیاز پایین میباشد.

Research_Output vs Score-4



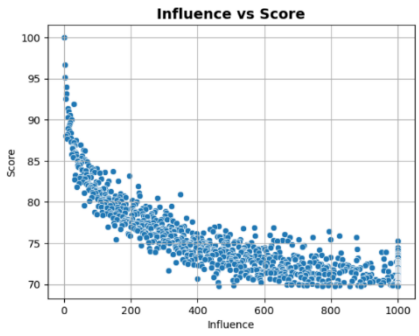
ویژگی خروجی تحقیق هر چقدر بیشتر، امتیاز پایین تر میرود.

Quality_Publications vs Score-5



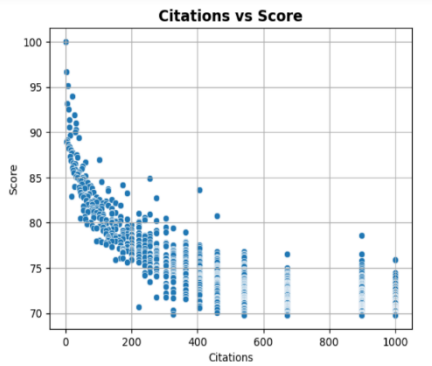
ویژگی انتشارات با کیفیت نیز همانند خروجی تحقیق با امتیاز رابطه عکس دارد.

Influence vs Score-6



ویژگی نفوذ نیز با امتیاز رابطه عکس دارد.

Citations vs Score-7

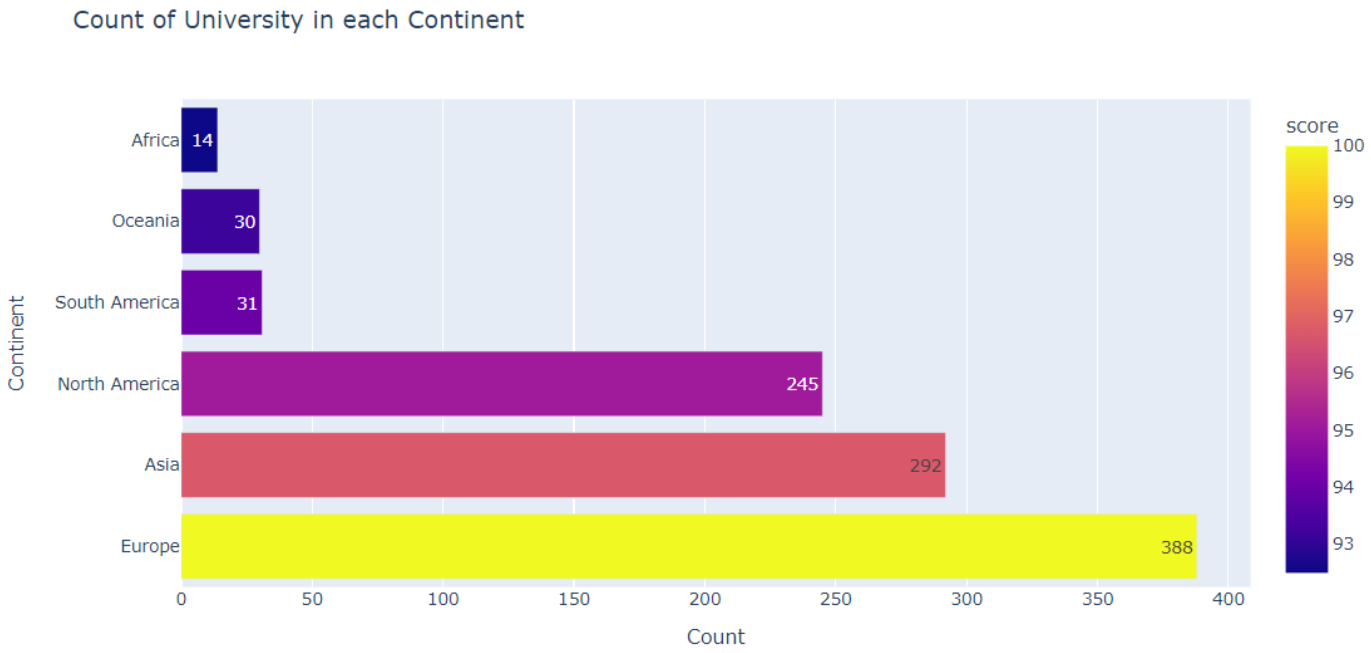


ویژگی استناد نیز با امتیاز رابطه عکس دارد.

نتیجه کلی از تاثیر این 7 ویژگی بر روی امتیاز:

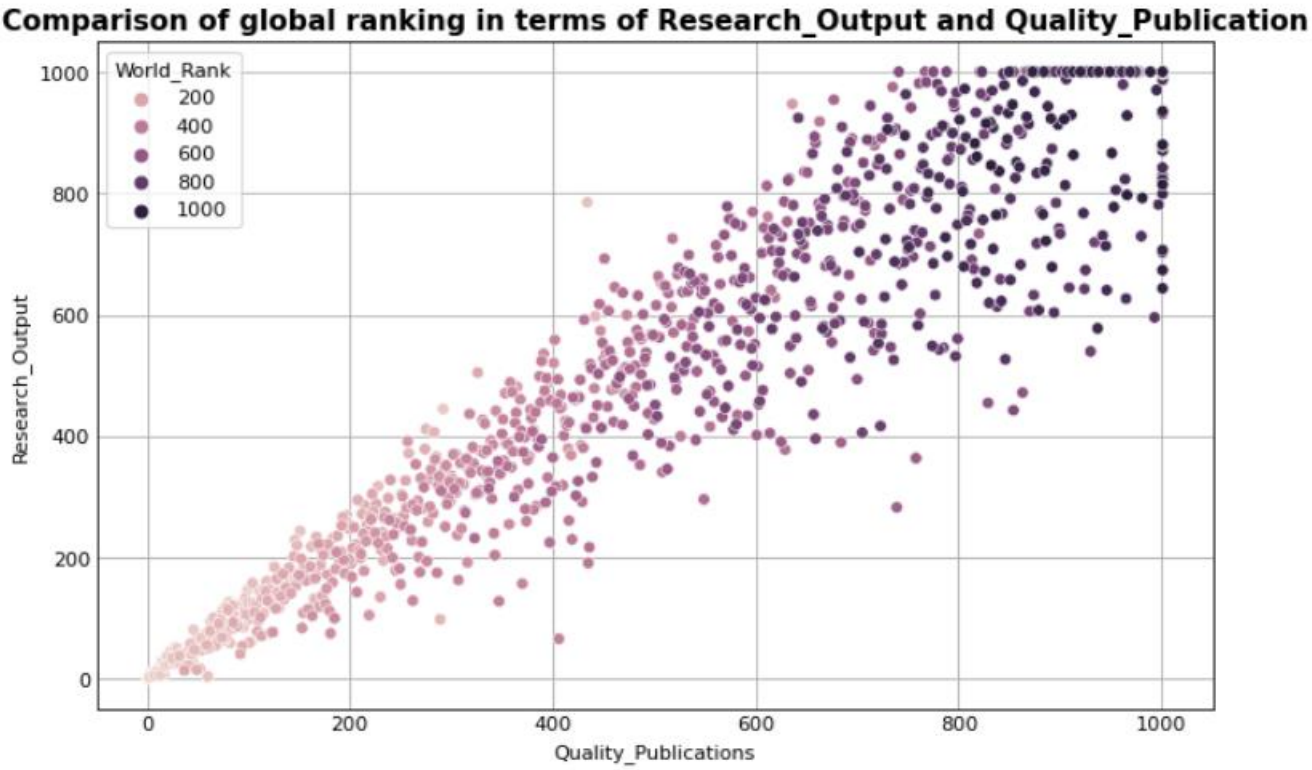
ویژگی های رتبه جهانی، رتبه ملی، انتشارات با کیفیت، نفوذ، خروجی تحقیق و استناد تقریبا با یکدیگر یکسان اند و تقریبا همبستگی یکسانی با امتیاز دارند. ولی ویژگی استخدام فارغ التحصیلان دارای پراکندگی و همبستگی زیاد با سایر ویژگی ها و امتیاز پایین است.

_تعداد دانشگاه های هر قاره



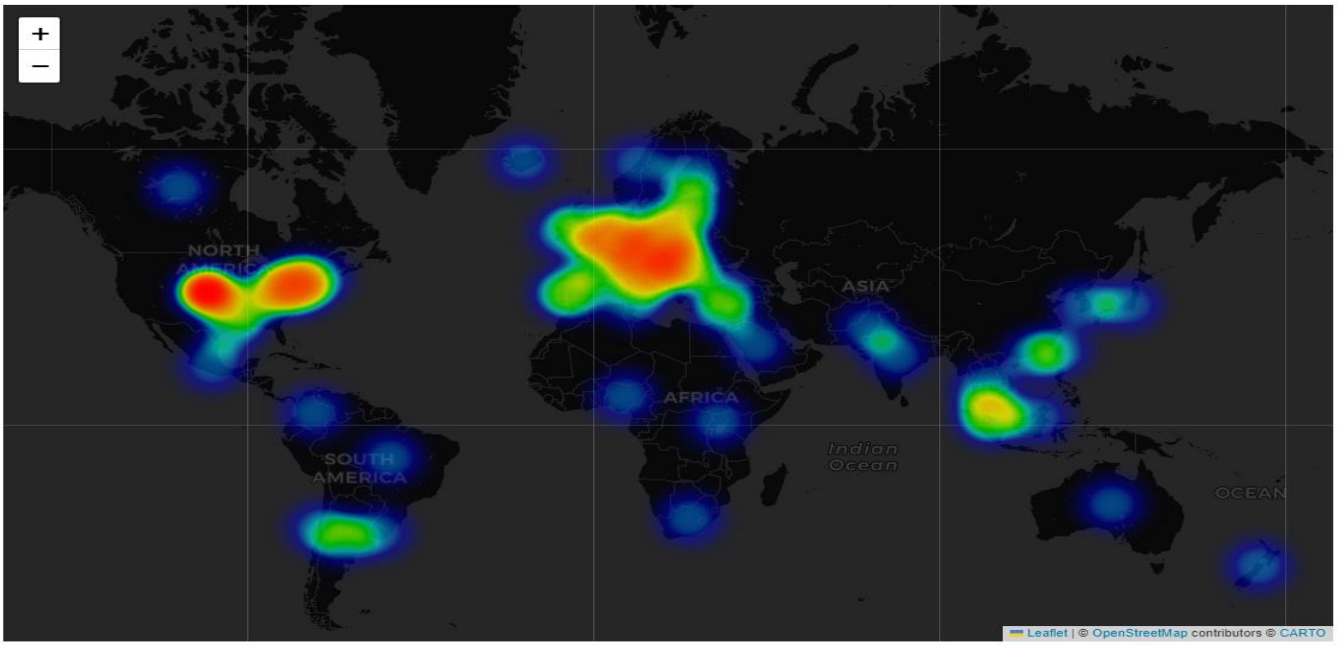
طبق این نمودار بیشتر دانشگاه ها در قاره اروپا قرار داردو بعد آن مربوط به آسیا و آمریکای شمالی میباشد.

_ مقایسه ی کیفیت انتشارات و خروجی تحقیق



طبق این نمودار ویژگی خروجی تحقیق و انتشارات با کیفیت با یکدیگر رابطهری خطی دارند و با افزایش یکی دیگری نیز افزایش می یابد و لی در حالت کلی با در نظر گرفتن نمودار های پراکندگی بالا این ویژگی ها با امتیاز کاملا رابطه عکس دارند.

_پراکندگی دانشگاه ها در قاره ها با نمودار هیت مپ



طبق نمودار هیت مپ نیز میتوان مشاهده کرد که تراکم و چگالی دانشگاه ها در این دیتااست در قاره اروپا بیشتر از سایر قاره ها می باشد.

_پراکندگی دانشگاه ها در قاره ها با لیبل گذاری قاره ها

در این نمودار نیز میتوان فهمید که بیشتر دانشگاه ها مربوط به کدام قاره میباشد که هر قاره با رنگ های خاص مشخص شده است.(اروپا = قرمز، آمریکا = نارنجی، آسیا = سبز، آفریقا = آبی، سایر = خاکستری)

میتوان مشاهده کرد که بیشترین مقدار مربوط به اروپا میباشد.