**מגישה: ציפי הורניק-לוריא**

**עבודת סיכום מס' 2 – EDA וניקוי נתונים**

**הערה כללית: אני לא הכנסתי את כל הפלטים/טבלאות/גרפים בקובץ סיכום זה**

**חלק 2 – בדיקת הנתונים**

בתחילה חילקתי את המשתנים ל -3 תתי data frame בהתאם לסוגי המשתנים – 1. משתנים רציפים 2. משתנים קטגוריאליים (עם יותר משתי קטגוריות) 3. משתנים דיכוטומיים

2.1 – תיאור הנתונים – סטטיסטיקה תיאורית

**Descriptive Continues variables:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | budget | popularity | runtime | revenue |
| valid\_N | 7.375000e+03 | 7375.0000000 | 7370.000000 | 2.993000e+03 |
| NAs | 0.000000e+00 | 0.0000000 | 5.000000 | 4.382000e+03 |
| means | 2.266666e+07 | 8.5397767 | 107.747218 | 6.687931e+07 |
| medi | 8.000000e+06 | 7.4558749 | 104.000000 | 1.692814e+07 |
| std | 3.698735e+07 | 12.1763844 | 21.412662 | 1.376564e+08 |
| CI.upper | 2.351095e+07 | 8.8177204 | 108.236159 | 7.181295e+07 |
| CI.mean | 2.266666e+07 | 8.5397767 | 107.747218 | 6.687931e+07 |
| CI.lower | 2.182237e+07 | 8.2618331 | 107.258278 | 6.194568e+07 |
| min | 0.000000e+00 | 0.0000010 | 0.000000 | 1.000000e+00 |
| max | 3.800000e+08 | 547.4882800 | 338.000000 | 1.519558e+09 |
| range | 3.800000e+08 | 547.4882790 | 338.000000 | 1.519558e+09 |
| percent5.5% | 0.000000e+00 | 0.6881493 | 83.000000 | 3.134560e+04 |
| percent95.95% | 1.000000e+08 | 17.6139879 | 145.550000 | 3.029503e+08 |
| skews | 2.886505e+00 | 19.9611460 | 1.160895 | 4.532977e+00 |
| kurt | 1.065443e+01 | 661.5187951 | 7.381437 | 2.765440e+01 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | keyword\_cnt | release\_year | sw\_collection | producers\_cnt |
| valid\_N | 7375.000000 | 7375.000000 | 7375.000000 | 6980.000000 |
| NAs | 0.000000 | 0.000000 | 0.000000 | 395.000000 |
| means | 7.264949 | 1999.677966 | 2.752271 | 2.907736 |
| medi | 6.000000 | 2004.000000 | 2.000000 | 2.000000 |
| std | 6.252185 | 15.373040 | 2.185598 | 2.143464 |
| CI.upper | 7.407664 | 2000.028878 | 2.802161 | 2.958030 |
| CI.mean | 7.264949 | 1999.677966 | 2.752271 | 2.907736 |
| CI.lower | 7.122234 | 1999.327054 | 2.702382 | 2.857443 |
| min | 0.000000 | 1918.000000 | 0.000000 | 1.000000 |
| max | 149.000000 | 2017.000000 | 26.000000 | 26.000000 |
| range | 149.000000 | 99.000000 | 26.000000 | 25.000000 |
| percent5.5% | 0.000000 | 1969.000000 | 0.000000 | 1.000000 |
| percent95.95% | 19.000000 | 2016.000000 | 7.000000 | 7.000000 |
| skews | 3.131014 | -1.689540 | 2.319777 | 2.500303 |
| kurt | 43.413735 | 3.724571 | 10.981312 | 12.057941 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | countries\_cnt | actor0\_movies\_cnt | actor0\_movies\_5y\_cnt |
| valid\_N | 7375.0000000 | 7375.000000 | 7375.0000000 |
| NAs | 0.0000000 | 0.000000 | 0.0000000 |
| means | 1.3350508 | 8.650441 | 3.3330169 |
| medi | 1.0000000 | 5.000000 | 3.0000000 |
| std | 0.7899121 | 9.993657 | 3.2986693 |
| CI.upper | 1.3530817 | 8.878560 | 3.4083139 |
| CI.mean | 1.3350508 | 8.650441 | 3.3330169 |
| CI.lower | 1.3170200 | 8.422321 | 3.2577200 |
| min | 0.0000000 | 0.000000 | 0.0000000 |
| max | 12.0000000 | 74.000000 | 19.0000000 |
| range | 12.0000000 | 74.000000 | 19.0000000 |
| percent5.5% | 1.0000000 | 0.000000 | 0.0000000 |
| percent95.95% | 3.0000000 | 29.000000 | 10.0000000 |
| skews | 3.0021910 | 1.639825 | 1.0359360 |
| kurt | 16.2316292 | 3.094175 | 0.6670633 |

|  |  |  |
| --- | --- | --- |
|  | actor1\_movies\_cnt | actor1\_movies\_5y\_cnt |
| valid\_N | 7375.000000 | 7375.000000 |
| NAs | 0.000000 | 0.000000 |
| means | 6.538169 | 2.723525 |
| medi | 3.000000 | 2.000000 |
| std | 8.583300 | 3.084940 |
| CI.upper | 6.734096 | 2.793944 |
| CI.mean | 6.538169 | 2.723525 |
| CI.lower | 6.342243 | 2.653107 |
| min | 0.000000 | 0.000000 |
| max | 79.000000 | 19.000000 |
| range | 79.000000 | 19.000000 |
| percent5.5% | 0.000000 | 0.000000 |
| percent95.95% | 24.000000 | 9.000000 |
| skews | 2.210898 | 1.366178 |
| kurt | 6.941230 | 1.780805 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | actor2\_movies\_cnt | actor2\_movies\_5y\_cnt | actor0\_prev\_revenue | actor1\_prev\_revenue |
| valid\_N | 7375.000000 | 7375.000000 | 5.060000e+03 | 4.574000e+03 |
| NAs | 0.000000 | 0.000000 | 2.315000e+03 | 2.801000e+03 |
| means | 5.748475 | 2.293288 | 2.228468e+08 | 2.050248e+08 |
| medi | 3.000000 | 1.000000 | 1.509680e+08 | 1.250639e+08 |
| std | 8.006172 | 2.816863 | 2.515196e+08 | 2.516415e+08 |
| CI.upper | 5.931227 | 2.357587 | 2.297786e+08 | 2.123193e+08 |
| CI.mean | 5.748475 | 2.293288 | 2.228468e+08 | 2.050248e+08 |
| CI.lower | 5.565722 | 2.228989 | 2.159149e+08 | 1.977303e+08 |
| min | 0.000000 | 0.000000 | 1.000000e+00 | 1.000000e+00 |
| max | 78.000000 | 19.000000 | 1.519558e+09 | 1.519558e+09 |
| range | 78.000000 | 19.000000 | 1.519558e+09 | 1.519558e+09 |
| percent5.5% | 0.000000 | 0.000000 | 3.484331e+06 | 3.335984e+06 |
| percent95.95% | 22.000000 | 8.000000 | 7.866360e+08 | 8.500000e+08 |
| skews | 2.378658 | 1.583630 | 2.171377e+00 | 2.295246e+00 |
| kurt | 8.046581 | 2.612479 | 5.935158e+00 | 6.145928e+00 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | actor2\_prev\_revenue | director\_movies\_cnt | director\_movies\_5y\_cnt | depart\_Art |
| valid\_N | 4.317000e+03 | 7375.000000 | 7375.0000000 | 7341.000000 |
| NAs | 3.058000e+03 | 0.000000 | 0.0000000 | 34.000000 |
| means | 1.958649e+08 | 1.957831 | 0.6508475 | 2.706716 |
| medi | 1.108795e+08 | 1.000000 | 0.0000000 | 1.000000 |
| std | 2.470005e+08 | 3.160345 | 0.9191578 | 4.118007 |
| CI.upper | 2.032350e+08 | 2.029970 | 0.6718286 | 2.800933 |
| CI.mean | 1.958649e+08 | 1.957831 | 0.6508475 | 2.706716 |
| CI.lower | 1.884947e+08 | 1.885691 | 0.6298663 | 2.612499 |
| min | 1.000000e+00 | 0.000000 | 0.0000000 | 0.000000 |
| max | 1.519558e+09 | 29.000000 | 6.0000000 | 36.000000 |
| range | 1.519558e+09 | 29.000000 | 6.0000000 | 36.000000 |
| percent5.5% | 2.747420e+06 | 0.000000 | 0.0000000 | 0.000000 |
| percent95.95% | 8.362972e+08 | 8.000000 | 2.0000000 | 12.000000 |
| skews | 2.272573e+00 | 3.004403 | 1.5764616 | 2.711877 |
| kurt | 5.896270e+00 | 12.836429 | 2.7491237 | 9.229593 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | depart\_Camera | depart\_Crew | depart\_Custom\_Mkup | depart\_Directing |
| valid\_N | 7341.000000 | 7341.000000 | 7341.000000 | 7341.000000 |
| NAs | 34.000000 | 34.000000 | 34.000000 | 34.000000 |
| means | 1.801662 | 2.387549 | 2.073696 | 1.665441 |
| medi | 1.000000 | 0.000000 | 1.000000 | 1.000000 |
| std | 2.588836 | 5.875461 | 3.553237 | 1.366409 |
| CI.upper | 1.860893 | 2.521976 | 2.154991 | 1.696703 |
| CI.mean | 1.801662 | 2.387549 | 2.073696 | 1.665441 |
| CI.lower | 1.742431 | 2.253123 | 1.992400 | 1.634178 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| max | 24.000000 | 119.000000 | 33.000000 | 30.000000 |
| range | 24.000000 | 119.000000 | 33.000000 | 30.000000 |
| percent5.5% | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| percent95.95% | 8.000000 | 14.000000 | 10.000000 | 4.000000 |
| skews | 2.739550 | 4.794105 | 2.551837 | 6.497113 |
| kurt | 9.182969 | 41.169472 | 7.839516 | 82.364233 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | depart\_Editing | depart\_Lighting | depart\_Production |
| valid\_N | 7341.000000 | 7341.0000000 | 7341.000000 |
| NAs | 34.000000 | 34.0000000 | 34.000000 |
| means | 1.516278 | 0.4262362 | 5.300776 |
| medi | 1.000000 | 0.0000000 | 4.000000 |
| std | 1.737599 | 1.2354498 | 5.269261 |
| CI.upper | 1.556033 | 0.4545024 | 5.421333 |
| CI.mean | 1.516278 | 0.4262362 | 5.300776 |
| CI.lower | 1.476523 | 0.3979700 | 5.180220 |
| min | 0.000000 | 0.0000000 | 0.000000 |
| max | 21.000000 | 18.0000000 | 38.000000 |
| range | 21.000000 | 18.0000000 | 38.000000 |
| percent5.5% | 0.000000 | 0.0000000 | 0.000000 |
| percent95.95% | 5.000000 | 3.0000000 | 16.000000 |
| skews | 2.301775 | 3.9512196 | 1.509471 |
| kurt | 7.931188 | 20.1313012 | 2.808198 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | depart\_Sound | depart\_Visual\_Effects | depart\_Writing |
| valid\_N | 7341.000000 | 7341.000000 | 7341.000000 |
| NAs | 34.000000 | 34.000000 | 34.000000 |
| means | 3.064296 | 1.290287 | 2.222041 |
| medi | 1.000000 | 0.000000 | 2.000000 |
| std | 4.515786 | 4.130573 | 1.627843 |
| CI.upper | 3.167614 | 1.384792 | 2.259284 |
| CI.mean | 3.064296 | 1.290287 | 2.222041 |
| CI.lower | 2.960979 | 1.195783 | 2.184797 |
| min | 0.000000 | 0.000000 | 0.000000 |
| max | 40.000000 | 89.000000 | 23.000000 |
| range | 40.000000 | 89.000000 | 23.000000 |
| percent5.5% | 0.000000 | 0.000000 | 1.000000 |
| percent95.95% | 13.000000 | 7.000000 | 5.000000 |
| skews | 2.215415 | 6.919387 | 2.803308 |
| kurt | 5.285101 | 79.011585 | 17.323074 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | depart\_Art\_female | depart\_Camera\_female | depart\_Crew\_female | depart\_Custom\_Mkup\_female |
| valid\_N | 7341.0000000 | 7.341000e+03 | 7.341000e+03 | 7341.0000000 |
| NAs | 34.0000000 | 3.400000e+01 | 3.400000e+01 | 34.0000000 |
| means | 0.2351178 | 2.043318e-02 | 6.443264e-02 | 0.3534941 |
| medi | 0.0000000 | 0.000000e+00 | 0.000000e+00 | 0.0000000 |
| std | 0.5580946 | 1.452870e-01 | 3.355672e-01 | 0.6515213 |
| CI.upper | 0.2478866 | 2.375724e-02 | 7.211017e-02 | 0.3684004 |
| CI.mean | 0.2351178 | 2.043318e-02 | 6.443264e-02 | 0.3534941 |
| CI.lower | 0.2223490 | 1.710912e-02 | 5.675511e-02 | 0.3385877 |
| min | 0.0000000 | 0.000000e+00 | 0.000000e+00 | 0.0000000 |
| max | 7.0000000 | 3.000000e+00 | 1.000000e+01 | 6.0000000 |
| range | 7.0000000 | 3.000000e+00 | 1.000000e+01 | 6.0000000 |
| percent5.5% | 0.0000000 | 0.000000e+00 | 0.000000e+00 | 0.0000000 |
| percent95.95% | 1.0000000 | 0.000000e+00 | 1.000000e+00 | 2.0000000 |
| skews | 3.0515403 | 7.570696e+00 | 9.947193e+00 | 2.2749406 |
| kurt | 13.5143049 | 6.660857e+01 | 1.783115e+02 | 6.9580737 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | depart\_Directing\_female | depart\_Editing\_female | depart\_Lighting\_female | depart\_Production\_female |
| valid\_N | 7341.0000000 | 7341.0000000 | 7.341000e+03 | 7341.000000 |
| NAs | 34.0000000 | 34.0000000 | 3.400000e+01 | 34.000000 |
| means | 0.1535213 | 0.1216456 | 6.811061e-04 | 0.939654 |
| medi | 0.0000000 | 0.0000000 | 0.000000e+00 | 0.000000 |
| std | 0.3944405 | 0.3579328 | 2.609090e-02 | 1.329413 |
| CI.upper | 0.1625458 | 0.1298348 | 1.278047e-03 | 0.970070 |
| CI.mean | 0.1535213 | 0.1216456 | 6.811061e-04 | 0.939654 |
| CI.lower | 0.1444968 | 0.1134563 | 8.416551e-05 | 0.909238 |
| min | 0.0000000 | 0.0000000 | 0.000000e+00 | 0.000000 |
| max | 4.0000000 | 5.0000000 | 1.000000e+00 | 12.000000 |
| range | 4.0000000 | 5.0000000 | 1.000000e+00 | 12.000000 |
| percent5.5% | 0.0000000 | 0.0000000 | 0.000000e+00 | 0.000000 |
| percent95.95% | 1.0000000 | 1.0000000 | 0.000000e+00 | 4.000000 |
| skews | 2.6869450 | 3.2337965 | 3.827012e+01 | 1.908707 |
| kurt | 8.1993269 | 13.5503879 | 1.462801e+03 | 5.098649 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | depart\_Sound\_female | depart\_Visual\_Effects\_female | depart\_Writing\_female |
| valid\_N | 7.341000e+03 | 7.341000e+03 | 7341.0000000 |
| NAs | 3.400000e+01 | 3.400000e+01 | 34.0000000 |
| means | 7.914453e-02 | 8.445716e-03 | 0.1242338 |
| medi | 0.000000e+00 | 0.000000e+00 | 0.0000000 |
| std | 3.164468e-01 | 9.444832e-02 | 0.3949095 |
| CI.upper | 8.638460e-02 | 1.060662e-02 | 0.1332690 |
| CI.mean | 7.914453e-02 | 8.445716e-03 | 0.1242338 |
| CI.lower | 7.190446e-02 | 6.284807e-03 | 0.1151985 |
| min | 0.000000e+00 | 0.000000e+00 | 0.0000000 |
| max | 6.000000e+00 | 2.000000e+00 | 5.0000000 |
| range | 6.000000e+00 | 2.000000e+00 | 5.0000000 |
| percent5.5% | 0.000000e+00 | 0.000000e+00 | 0.0000000 |
| percent95.95% | 1.000000e+00 | 0.000000e+00 | 1.0000000 |
| skews | 5.110545e+00 | 1.169553e+01 | 3.7497033 |
| kurt | 3.879563e+01 | 1.468359e+02 | 17.6772230 |

**Descriptive Categorical variables**

$original\_language

|  |  |
| --- | --- |
| validN | 7375 |
| NAs | 0 |
| mods | en |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **category** | **n** | **category** | **n** | **category** | **n** | **category** | **n** | **category** | **n** | **category** | **n** |
| af | 1 | cn | 41 | fi | 3 | ja | 90 | no | 5 | te | 9 |
| ar | 1 | cs | 3 | fr | 199 | ka | 1 | pl | 5 | th | 5 |
| bm | 2 | da | 17 | he | 6 | kn | 1 | pt | 13 | tr | 9 |
| bn | 3 | de | 49 | hi | 118 | ko | 48 | ro | 8 | ur | 2 |
|  |  | el | 3 | hu | 4 | ml | 12 | ru | 105 | vi | 1 |
|  |  | en | 6336 | id | 3 | mr | 1 | sr | 3 | xx | 2 |
|  |  | es | 95 | is | 1 | nb | 1 | sv | 20 | zh | 46 |

$runtime\_cat

|  |  |
| --- | --- |
| validN | 7375 |
| NAs | 0 |
| mods | Medium |

|  |  |
| --- | --- |
| **category** | **n** |
| Large | 1769 |
| Medium | 3884 |
| Short | 1722 |

$release\_year

|  |  |
| --- | --- |
| validN | 7375 |
| NAs | 0 |
| mods | 2013 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Release year | 1918 | 1921 | 1922 |  |  |  |
| validN | 1 | 1 | 1 |  |  |  |
| Release year | 1923 | 1924 | 1925 | 1926 | 1927 | 1928 |
| validN | 1 | 1 | 6 | 2 | 3 | 5 |
| Release year | 1929 | 1930 | 1931 | 1932 | 1933 | 1934 |
| validN |  | 1 | 2 | 4 | 3 | 6 |
| Release year | 1935 | 1936 | 1937 | 1938 | 1939 | 1940 |
| validN | 5 | 6 | 2 | 3 | 6 | 5 |
| Release year | 1941 | 1942 | 1943 | 1944 | 1945 | 1946 |
| validN | 6 | 6 | 2 | 7 | 6 | 4 |
| Release year | 1947 | 1948 | 1949 | 1950 | 1951 | 1952 |
| validN | 2 | 9 | 5 | 6 | 9 | 7 |
| Release year | 1953 | 1954 | 1955 | 1956 | 1957 | 1958 |
| validN | 11 | 14 | 10 | 13 | 7 | 8 |
| Release year | 1959 | 1960 | 1961 | 1962 | 1963 | 1964 |
| validN | 10 | 14 | 16 | 15 | 15 | 14 |
| Release year | 1965 | 1966 | 1967 | 1968 | 1969 | 1970 |
| validN | 12 | 15 | 24 | 23 | 15 | 23 |
| Release year | 1971 | 1972 | 1973 | 1974 | 1975 | 1976 |
| validN | 20 | 16 | 23 | 19 | 25 | 22 |
| Release year | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 |
| validN | 29 | 43 | 40 | 59 | 65 | 67 |
| Release year | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 |
| validN | 79 | 78 | 94 | 102 | 112 | 132 |
| Release year | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 |
| validN | 102 | 104 | 116 | 116 | 138 | 123 |
| Release year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 |
| validN | 134 | 138 | 158 | 152 | 158 | 156 |
| Release year | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
| validN | 175 | 197 | 178 | 208 | 234 | 280 |
| Release year | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
| validN | 257 | 265 | 284 | 304 | 311 | 304 |
| Release year | 2013 | 2014 | 2015 | 2016 | 2017 |  |
| validN | 334 | 319 | 312 | 300 | 98 |  |

$release\_month

|  |  |
| --- | --- |
| validN | 7375 |
| NAs | 0 |
| mods | 9 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Release month | 1 | 2 | 3 | 4 | 5 |  |  |  |
| validN | 510 | 518 | 561 | 557 | 583 |  |  |  |
| Release month | 6 | 7 | 8 | 9 | 10 | 11 | 12 |  |
| validN | 594 | 564 | 657 | 902 | 695 | 540 | 694 |  |

$release\_day

|  |  |
| --- | --- |
| validN | 7375 |
| NAs | 0 |
| mods | 1 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Release day | 1 | 2 | 3 | 4 | 5 |  |  |  |
| validN | 375 | 215 | 198 | 191 | 216 |  |  |  |
| Release day | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| validN | 261 | 255 | 229 | 262 | 261 | 262 | 275 | 267 |
| Release day | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| validN | 262 | 276 | 257 | 245 | 270 | 229 | 232 | 258 |
| Release day | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| validN | 262 | 220 | 231 | 289 | 208 | 195 | 197 | 168 |
| Release day | 30 | 31 |  |  |  |  |  |  |
| validN | 193 | 116 |  |  |  |  |  |  |

**Descriptive Dichotomous Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sw\_lang\_en | sw\_web\_presence | sw\_has\_poster | sw\_tagline |
| validN | 7375 | 7375 | 7375 | 7375 |
| NAs | 0 | 0 | 0 | 0 |
| mods | 1 | 0 | 1 | 1 |
| Value 0 | 1039 | 5012 | 2 | 1442 |
| Value 1 | 6336 | 2363 | 7373 | 5933 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | high\_release\_month | lang\_US | lang\_FR | lang\_RU | lang\_ES | lang\_JA |
| validN | 7375 | 7375 | 7375 | 7375 | 7375 | 7375 |
| NAs | 0 | 0 | 0 | 0 | 0 | 0 |
| mods | 0 | 1 | 0 | 0 | 0 | 0 |
| Value 0 | 4427 | 933 | 6664 | 7017 | 6816 | 7171 |
| Value 1 | 2948 | 6442 | 711 | 358 | 559 | 204 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | sw\_female\_actor0 | sw\_female\_actor1 | sw\_female\_actor2 |
| validN | 7054 | 7219 | 7213 |
| NAs | 321 | 156 | 162 |
| mods | NA | NA | NA |
| Value 0 | 5413 | 4491 | 4764 |
| Value 1 | 1641 | 2728 | 2449 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sw\_male\_actor0 | sw\_male\_actor1 | sw\_male\_actor2 | genre\_adventure |
| validN | 7054 | 7219 | 7213 | 7375 |
| NAs | 321 | 156 | 162 | 0 |
| mods | NA | NA | NA | 0 |
| Value 0 | 2142 | 3472 | 3304 | 6259 |
| Value 1 | 4912 | 3747 | 3909 | 1116 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | genre\_fantasy | genre\_animation | genre\_drama | genre\_horror |
| validN | 7375 | 7375 | 7375 | 7375 |
| NAs | 0 | 0 | 0 | 0 |
| mods | 0 | 0 | 0 | 0 |
| Value 0 | 6747 | 6993 | 3699 | 6640 |
| Value 1 | 628 | 382 | 3676 | 735 |

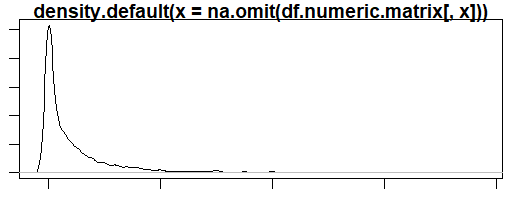
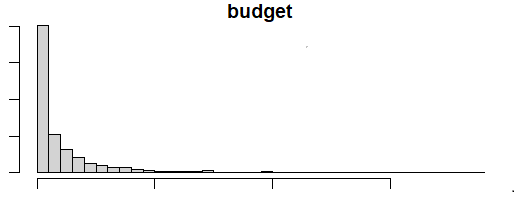
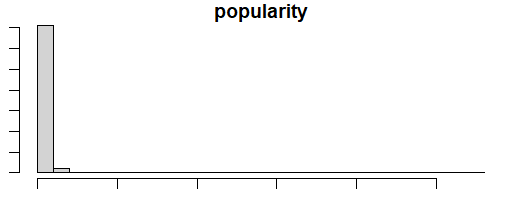
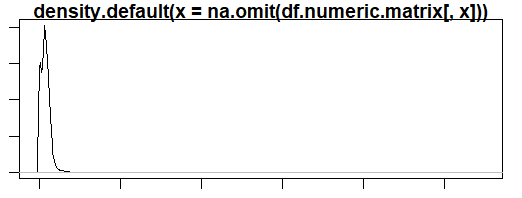
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | genre\_action | genre\_comedy | genre\_history | genre\_western |
| validN | 7375 | 7375 | 7375 | 7375 |
| NAs | 0 | 0 | 0 | 0 |
| mods | 0 | 0 | 0 | 0 |
| Value 0 | 5640 | 4770 | 7080 | 7258 |
| Value 1 | 1735 | 2605 | 295 | 117 |

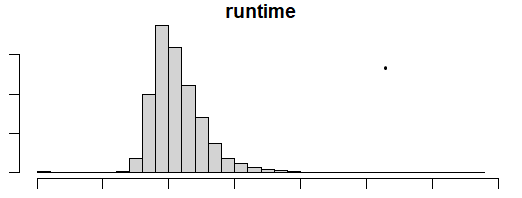
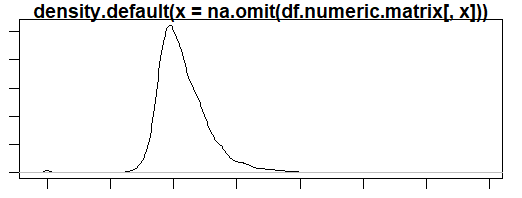
|  |  |  |  |
| --- | --- | --- | --- |
|  | genre\_thriller | genre\_crime | genre\_documentary |
| validN | 7375 | 7375 | 7375 |
| NAs | 0 | 0 | 0 |
| mods | 0 | 0 | 0 |
| Value 0 | 5506 | 6291 | 7154 |
| Value 1 | 1869 | 1084 | 221 |

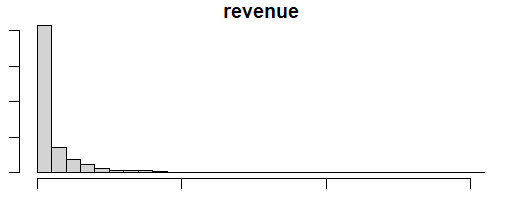
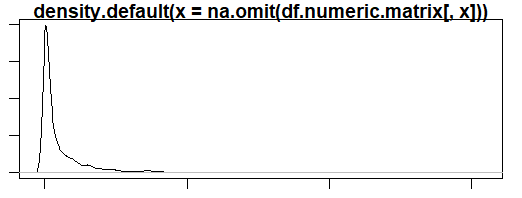
|  |  |  |  |
| --- | --- | --- | --- |
|  | genre\_science\_fiction | genre\_mystery | genre\_music |
| validN | 7375 | 7375 | 7375 |
| NAs | 0 | 0 | 0 |
| mods | 0 | 0 | 0 |
| Value 0 | 6631 | 6825 | 7108 |
| Value 1 | 744 | 550 | 267 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | genre\_romance | genre\_family | genre\_war | genre\_foreign |
| validN | 7375 | 7375 | 7375 | 7375 |
| NAs | 0 | 0 | 0 | 0 |
| mods | 0 | 0 | 0 | 0 |
| Value 0 | 5940 | 6700 | 7132 | 7291 |
| Value 1 | 1435 | 675 | 243 | 84 |

2.2 – ייצור גרפים המתארים את ההתנהגות של כל משתנה:

דוגמאות לגרפים של משתנים רציפים – היסטוגרמות + desity:





דוגמאות לבדיקת נורמאליות של כל המשתנים הרציפים בעזרת Two-sample Kolmogorov-Smirnov test

budget

D = 0.72529, p-value < 2.2e-16

popularity

D = 0.83653, p-value < 2.2e-16

runtime

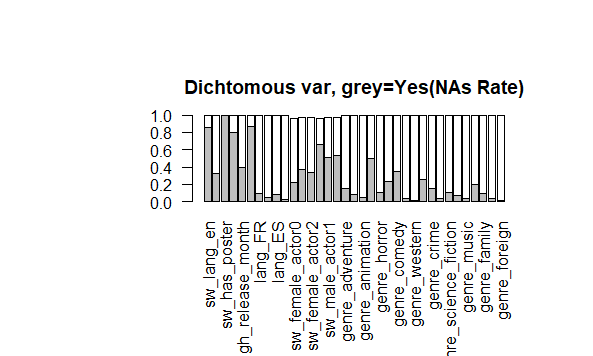
D = 0.99742, p-value < 2.2e-16

revenue

D = 0.99733, p-value < 2.2e-16

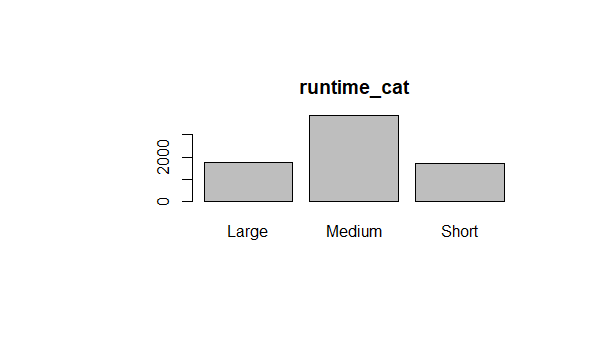
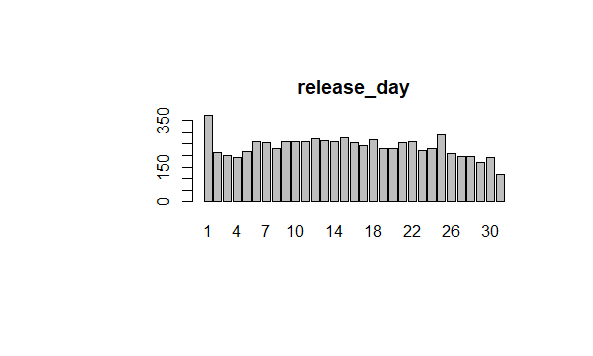
**המסקנה מכל הסטטיסטיקה התיאורית – עבור המשתנים הרציפים היא – שכולם ללא יוצא מן הכלל לא מתפלגים נורמאלית!**

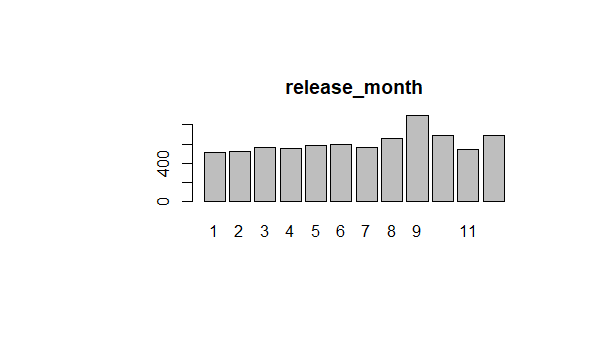
תיאור גרפי של כל המשתנים הדיכוטומיים – גרף עמודות



מהגרף ומהסטטיסטיקה התיאורית של המשתנים הדיכוטומיים, ניתן ללמוד שישנם משתנים כגון: sw\_has\_poster, lang\_US – שאין בהם שונות – כמעט לכל הסרטים היה פוסטר וכמעט לכל הסרטים היה דיבוב לאנגלית (לכן משתנים אילו חסרי משמעות כמסבירי שונות).

תיאור גרפי של המשתנים הקטגוריאליים:





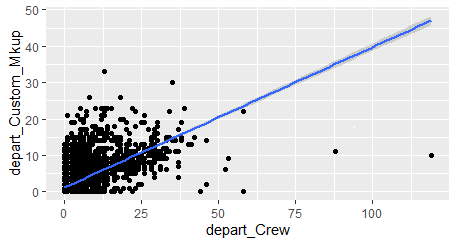
מהגרפים השונים ומהסטטיסטיקה התיאורית ניתן ללמוד שאורך רוב הסרטים הוא בינוני, היום הפופולרי לפרסום סרט הוא היום הראשון של החודש. בחודשים 9-12 יוצאים רוב הסרטים.

רוב השחקנים הראשיים הם גברים, ככל שעוברות השנים יש יותר סרטים

2.3 – בחרתי scatter plot – כדי לתאר את כל הקורלציות בין כל המשתנים הרציפים.

מכיוון שכאמור כל המשתנים הרציפים לא נמצאו כמתפלגים נורמאלית – ביצעתי מטריצת קורלציות של ספירמן בין כל המשתנים הרציפים והדיכוטומיים.

דוגמא ל scatter plot



דוגמא לקורלציה של ספירמן:

ערך R

keyword\_cnt

budget 0.283028503

popularity 0.431473557

runtime 0.143200997

revenue 0.332267153

sw\_lang\_en 0.168981170

sw\_web\_presence 0.146961593

sw\_tagline 0.281271088

keyword\_cnt 1.000000000

high\_release\_month 0.005350436

release\_day 0.020259794

sw\_collection 0.244501581

producers\_cnt 0.169521779

countries\_cnt 0.128883583

ערך המובהקות – p value

keyword\_cnt

budget 0.0000

popularity 0.0000

runtime 0.0000

revenue 0.0000

sw\_lang\_en 0.0000

sw\_web\_presence .0000

sw\_tagline 0.0000

keyword\_cnt 0.0000

high\_release\_month 0.6459

release\_day 0.0819

sw\_collection 0.0000

producers\_cnt 0.0000

countries\_cnt 0.0000

מספר הזוגות - N

keyword\_cnt

budget 7375

popularity 7375

runtime 7370

revenue 2993

sw\_lang\_en 7375

sw\_web\_presence 7375

sw\_tagline 7375

keyword\_cnt 7375

high\_release\_month 7375

release\_day 7375

sw\_collection 7375

producers\_cnt 6980

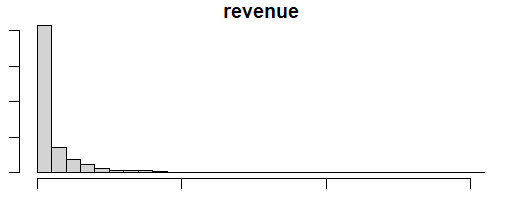
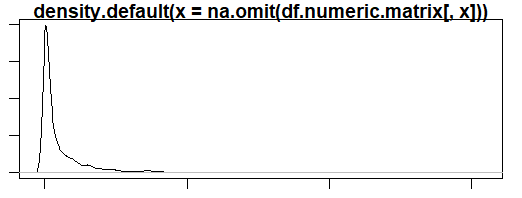
countries\_cnt 7375

מכיוון שישנם משתנים עם ערכים חסרים השתמשתי בפקודה לקורלציה שמאפשרת pairwise

2.4 משתנה המטרה – REVENUE

בסעיף 2.1 תוארו כל המשתנים הרציפים ובינהם משתנה המטרה revenue:

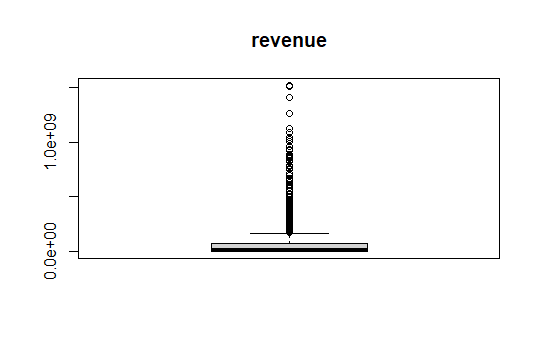
|  |  |
| --- | --- |
|  | revenue |
| valid\_N | 2.993000e+03 |
| NAs | 4.382000e+03 |
| means | 6.687931e+07 |
| medi | 1.692814e+07 |
| std | 1.376564e+08 |
| CI.upper | 7.181295e+07 |
| CI.mean | 6.687931e+07 |
| CI.lower | 6.194568e+07 |
| min | 1.000000e+00 |
| max | 1.519558e+09 |
| range | 1.519558e+09 |
| percent5.5% | 3.134560e+04 |
| percent95.95% | 3.029503e+08 |
| skews | 4.532977e+00 |
| kurt | 2.765440e+01 |

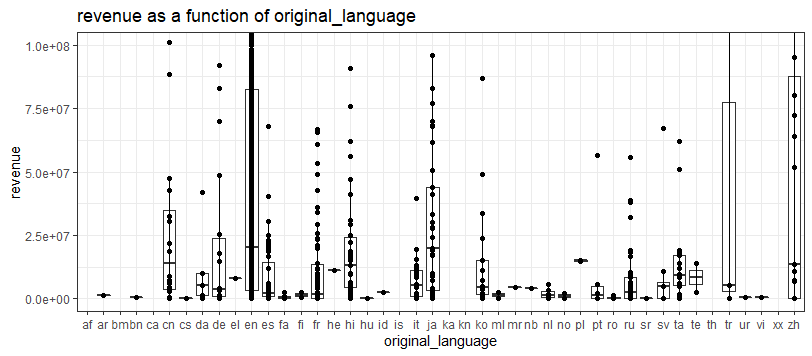
בדיקת נורמאליות של משתנה המטרה בעזרת Two-sample Kolmogorov-Smirnov test:

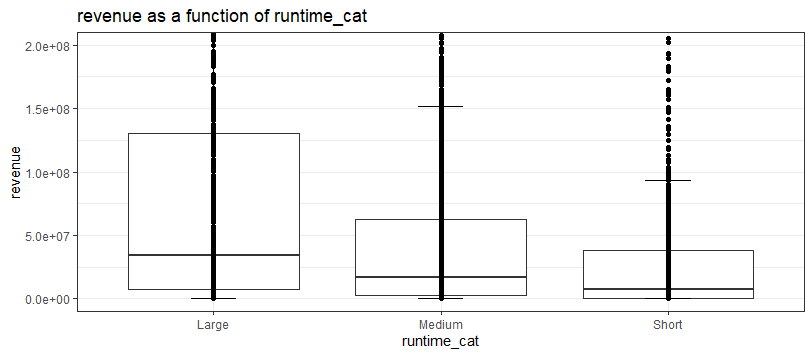
D = 0.99733, p-value < 2.2e-16

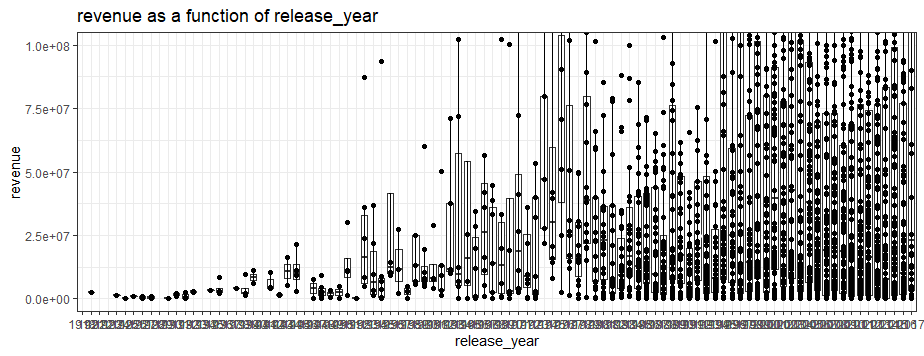
בסעיף 2.5 ביצעתי עבור כל המשתנים הרציפים – כולל משתנה המטרה בדיקה של ערכי קיצון בעזרת גרף box plot, מצ"ב הגרף:

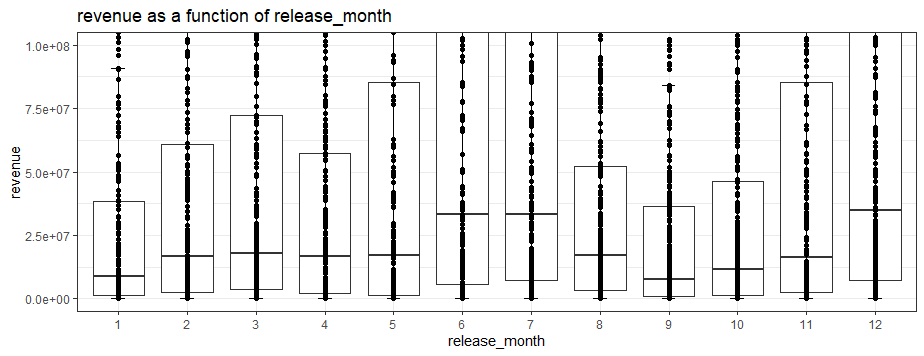


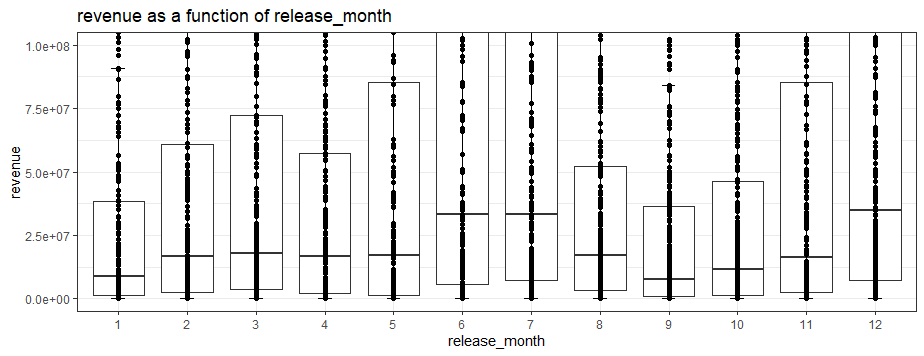
בנוסף, לבחינת משתנה המטרה שנעשתה בסעיף 2.1 ו- 2.5 – בחנתי את הקשר בין משתנה המטרה לחמשת המשתנים הקטגוריאליים בעזרת חמישה גרפים box plot.





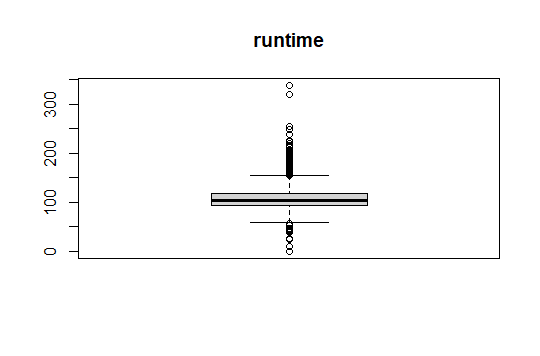
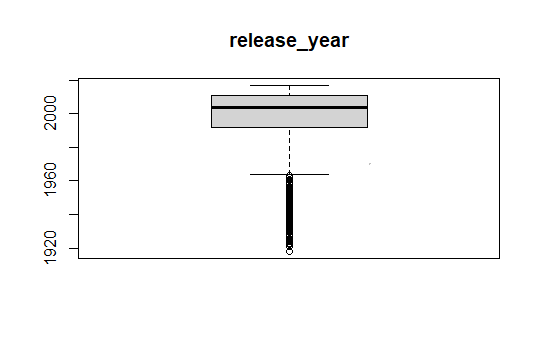
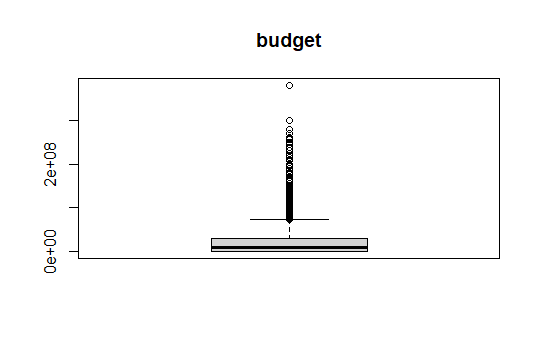






**מסקנה: רק לכ-3000 סרטים מתוך 7375 ישנו ערך במשתנה המטרה – הרווחיות - משתנה זה אינו מתפלג נורמאלית, ההתפלגות היא עם זנב ימני – מעט סרטים הם עם רווחיות גבוהה מאד, זהו משתנה עם הרבה ערכי קיצון.**

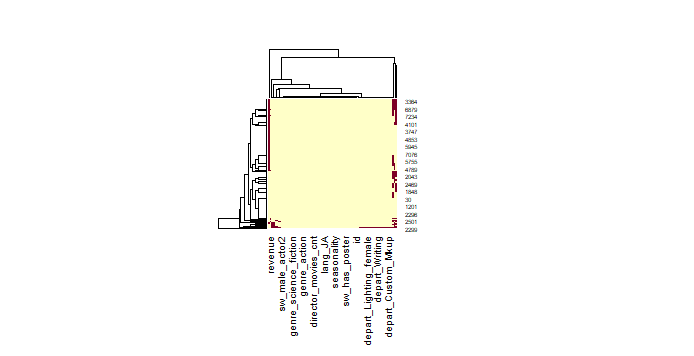
2.5 – גרפים הבוחנים ערכי קיצון – box plot – בתוכנית ביצעתי עבור כל המשתנים הרציפים בדיקה לערכי קיצון. פה מובאות מספר דוגמאות בלבד:



**מסקנה:** רוב המשתנים הם עם ערכי קיצון – בערכים הגבוהים – התפלגות עם זנב ימני – כאמור בתוכנית הרצתי את כל המשתנים הרציפים. במסמך זה ישנה דוגמא לערכי קיצון גבוהים – זנב ימני במשתנה budget וגם במשתנה revenue, משתנה release year הוא דוגמא למשתנה עם ערכי קיצון נמוכים – בשנים הראשונות ישנם פחות סרטים פר-שנה לעומת השנים האחרונות – זו התפלגות עם זנב שמאלי. משתנה runtime הוא דוגמא למשתנה עם ערכי קיצון גבוהים ונמוכים – דו זנבי – עם נטייה קלה לזנב ימני, יש יותר סרטים עם ערכי קיצון שזמן הסרט ארוך לעומת ערכי קיצון של סרטים קצרים.

**2.6 – ערכים חסרים חושבו בסעיף 2.1 – המשתנה שיש לו הכי הרבה ערכים חסרים הוא revenue. המשתנים שבחנו את הרווחיות ההיסטורית של השחקן הראשי ושל שני שחקני המשנה (**actor0\_prev\_revenue actor1\_prev\_revenue;actor1\_prev\_revenue; **)– נבנו על בסיס מסד נתונים זה, של רווחיות הסרטים – לכן, זה איננו מפתיע שמשתנים אילו גם הם עם ערכים חסרים רבים.**

2.7 –יצרתי מטריצה הבוחנת את הימצאות הערכים החסרים - missing.matrix – בה לכל משתנה היו ערכים 0 (לא ערך חסר) 1 (ערך חסר) – על בסיס משתנים אלו נוצר הגרף heatmap המוצג :



**3 – ניקוי הנתונים**

**3.1 – נתוני קיצון**

1. על בסיס הנוסחה: Q[1,2]±1.5\*iqr – הושארו בכל אחד מהמשתנים הרציפים (שנמצאו בהם ערכי קיצון )- רק ערכים שאינם ערכי קיצון. נעשו גרפים עבור המשתנים עם ערכי הקיצון ועם אותם המשתנים לאחר ניקוי ערכי הקיצון. מהתבוננות והשוואה בין הגרפים לא נראה שישנם הבדלים משמעותיים בהתפלגויות לאחר הורדת ערכי הקיצון.

2. יוצרו גרפים scatter – בין המשתנים ללא ערכי הקיצון ומשתנה המטרה. המשתנים שבהם נראה הבדל הם budget; popularity, keyword\_cnt – ובכמה ממשתני ה , depart\_ variables.

3. בחלק מהמשתנים נמצאו הבדלים מובהקים בקורלציות לאחר הסרת ערכי הקיצון . יחד עם זאת, הבדלים אלו היו קטנים וללא השפעה מהותית על התוצאות.

4. לאור האמור בסעיף 3, לא התבצעה מחיקה של ערכי הקיצון.

3.2 – כפי שכתבתי בסעיף 2.6 –המשתנה שיש לו הכי הרבה ערכים חסרים הוא revenue. וכן המשתנים שבחנו את הרווחיות ההיסטורית של השחקן הראשי ושל שני שחקני המשנה (actor0\_prev\_revenue actor1\_prev\_revenue; actor1\_prev\_revenue; )– משתנים אלו נבנו על ידי ובוחנים את הרווחיות ההיסטורית (הרווחיות של הסרטים הקודמים שנמצאים במסד הנתונים הנוכחי) של כל אחד משחקנים אילו (הרווחיות היא כידוע המשתנה revenue שנמצא במסד נתונים זה ושהוא בעל הרבה ערכים חסרים) , ההסתמכות של משתנים אילו על משתנה במסד נתונים זה שהוא בעל ערכים חסרים יוצרת את המנגנון של הערכים החסרים – בקובץ זה, שכולם נשענים על revenue.

טכניקת ה- imputation - שהשתמשתי בה היא Multivariate imputation by chained equations (MICE) – זו טכניקה שבוחרת "לבד" איך למלא את הערכים החסרים.

חלק 4 – אני מצטערת אבל לא הספקתי להריץ את סעיף 2 על הקובץ הנקי