In [1]:
```
from google.colab import drive

drive.mount("/content/gdrive")
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
```

In [2]:
```
%cd "/content/gdrive/MyDrive/Bloque IA/Estadistica/Entregable"
!ls
```

```
/content/gdrive/MyDrive/Bloque IA/Estadistica/Entregable
Entregable.ipynb  us2022q2a.gsheet  usfirms2022.gsheet
us2022q2a.csv     usfirms2022.csv
```

In [3]:
```
import plotly.express as px
import pandas as pd
import numpy as np
```

In [4]:
```
# The file has no names for columns.
df1 = pd.read_csv('us2022q2a.csv')
df2 = pd.read_csv('usfirms2022.csv')
```

DROPEAMOS LAS COLUMNAS DE FISCALMONTH, YEAR Y CTO, YA QUE ESTOS DATOS SE PUEDEN VER RESUMIDOS EN LA COLUMNA Q. EL RESTO DE DATOS SE MATIENEN YA QUE CON ESTOS CALCULAREMOS EL BOOK VALUE, MARKET VALUE Y OPERATING PROFIT MARGIN

In [5]:
```
df1 = df1.drop(['fiscalmonth', 'year', 'cto'], axis=1)
```

DROPEAMOS LAS COLUMNAS N, COUNTRY OF ORIGIN Y TYPE OF ASSET, YA QUE ESTOOS SON DATOS IGUALES PARA TODOS LOS REGUSTROS Y DEBIDO A ESTO NO APORTAN AL MODELO

In [6]:
```
df2 = df2.drop(['N', 'Country\nof Origin', 'Type of Asset'], axis=1)
```

2.2.1.1

Show how many firms by industry there are in the sample

OBTENEMOS EL TOTAL DE FIRMAS ENFOCADAS A CADA SECTOR

In [7]:
```
a = df2['Sector NAICS\nlevel 1'].value_counts()

a
```

Out[7]:
```
Manufacturing                                                  1567
Finance and Insurance                                           703
Information                                                     263
Retail Trade                                                    152
Professional, Scientific, and Technical Services                145
Administrative and Support and Waste Management and Remediation Services    133
Mining, Quarrying, and Oil and Gas Extraction                   104
Wholesale Trade                                                  79
Utilities                                                        77
Transportation and Warehousing                                   69
Accommodation and Food Services                                  69
Real Estate and Rental and Leasing                               68
Health Care and Social Assistance                                64
Construction                                                     45
Arts, Entertainment, and Recreation                              22
Other Services (except Public Administration)                    16
Agriculture, Forestry, Fishing and Hunting                       16
Educational Services                                             14
-                                                                 2
Name: Sector NAICS\nlevel 1, dtype: int64
```

In [8]:
```
df_merge = df1.merge(df2, left_on='firm', right_on='Ticker')
df_merge
```

Out[8]:

	firm	q	revenue	cogs	sgae	otheropexp	extraincome	finexp	incometax	totalassets	...	originalprice	sharesoutstanding	Ticker	Nam
0	A	2000q1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	104.0000	452000.000	A	Agilen Technologie Ir
1	A	2000q2	2485000.0	1261000.0	1.010000e+06	0.0	42000.000000	0.000	90000.0	7321000.000	...	73.7500	452271.967	A	Agilen Technologie Ir
2	A	2000q3	2670000.0	1369000.0	1.091000e+06	0.0	28000.000000	0.000	83000.0	7827000.000	...	48.9375	453014.579	A	Agilen Technologie Ir
3	A	2000q4	3372000.0	1732000.0	1.182000e+06	0.0	10000.000000	0.000	163000.0	8425000.000	...	54.7500	456366.381	A	Agilen Technologie Ir
4	A	2001q1	2841000.0	1449000.0	1.113000e+06	0.0	-6000.000000	0.000	119000.0	9208000.000	...	30.7300	456769.737	A	Agilen Technologie Ir
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
323811	ZYNE	2021q2	0.0	0.0	9.838494e+03	0.0	-117.528220	-5.943	0.0	98195.904	...	5.2900	41251.537	ZYNE	Zynerb Pharmaceutical Ir
323812	ZYNE	2021q3	0.0	0.0	1.021065e+04	0.0	-376.636750	-5.038	0.0	89996.170	...	4.2400	41251.537	ZYNE	Zynerb Pharmaceutical Ir
323813	ZYNE	2021q4	0.0	0.0	8.836436e+03	0.0	16.937906	-4.433	0.0	81171.507	...	2.8800	41217.537	ZYNE	Zynerb Pharmaceutical Ir
323814	ZYNE	2022q1	0.0	0.0	8.903915e+03	0.0	317.252110	-96.044	0.0	74381.029	...	2.0500	42447.037	ZYNE	Zynerb Pharmaceutical Ir

CREAMOS UN NUEVO DATAFAME DONDE SOLO TENEMOS LOS DATOS DEL ULTIMO TREIMESTRE DE 2022

In [9]:
```
df_mask = df_merge['q'] == '2022q2'
df3 = df_merge[df_mask]
```

CREAMOS NUEVOS CAMPOS EN EL DATAFRAME, DONDE GUARDAMOS EL BOOK VALUE, MARKET VALUE Y OPERATING PROFIT MARGIN DE CADA EMPRESA

In [10]:
```
df3['Book'] = df3['totalassets'] - df3['totalliabilities']
df3['Market'] = df3['originalprice'] * df3['sharesoutstanding']
df3['ebit'] = df3['revenue'] - df3['cogs'] - df3['sgae'] - df3['otheropexp']
df3['OPM'] = df3['ebit']  / df3['revenue']
```

In [11]:
```
df3
```

Out[11]:

	firm	q	revenue	cogs	sgae	otheropexp	extraincome	finexp	incometax	totalassets	...	Class	Sector NAICS\nlevel 1	Exchange / Src	Sector\nEconom
89	A	2022q2	1607000.0	746000.0	5.010000e+05	0.0	-7000.00000	20000.000	59000.0	1.045500e+07	...	Com	Manufacturing	NYSE	Electric Ele
179	AA	2022q2	3644000.0	2767000.0	2.200000e+05	-75000.0	81000.00000	30000.000	234000.0	1.570900e+07	...	Com	Manufacturing	NYSE	Basic & Fab M
269	AAIC	2022q2	10900.0	6374.0	0.000000e+00	0.0	-3417.00000	0.000	802.0	1.084755e+06	...	Com A	Finance and Insurance	NYSE	F
359	AAL	2022q2	13422000.0	0.0	1.240500e+07	0.0	25000.00000	439000.000	127000.0	6.796300e+07	...	Com	Transportation and Warehousing	NASDAQ	Transportat
449	AAME	2022q2	44669.0	0.0	4.678400e+04	0.0	0.00000	0.000	-436.0	3.792740e+05	...	Com	Finance and Insurance	NASDAQ	Finance and Insur
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
323455	ZVIA	2022q2	45542.0	28168.0	2.407400e+04	8043.0	3662.00000	0.000	9.0	1.127380e+05	...	Com A	Manufacturing	NYSE	Food & Beve
323545	ZVO	2022q2	51380.0	34995.0	2.610800e+04	-8882.0	-3824.00000	0.000	8.0	8.543300e+04	...	Com	Educational Services	NASDAQ	C
323635	ZWS	2022q2	284200.0	170400.0	6.000000e+04	300.0	-600.00000	5200.000	11300.0	1.176300e+06	...	Com	Manufacturing	NYSE	Industrial Ma
323725	ZY	2022q2	2634.0	9732.0	5.863800e+04	40460.0	-885.00000	9376.000	11.0	4.709680e+05	...	Com	Professional, Scientific, and Technical Services	NASDAQ	C

For each industry (and for all industries), what can you say about the typical firm size in terms of market value and book value? How much these variables change within each industry? How firm size (in market value) is distributed?

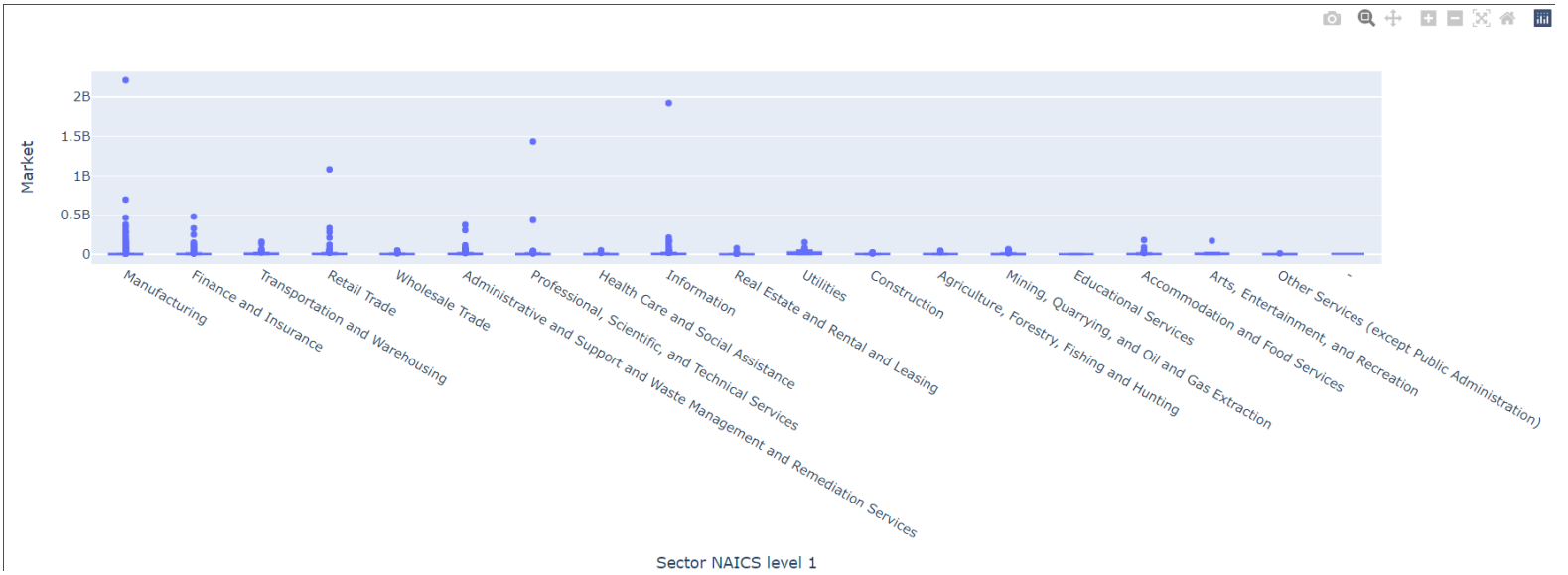
In [12]:
```
df_description = df3.groupby('Sector NAICS\nlevel 1')['Book'].mean().to_frame()
df_description['Book median'] = df3.groupby('Sector NAICS\nlevel 1')['Book'].median()
df_description['Market mean'] = df3.groupby('Sector NAICS\nlevel 1')['Market'].mean()
df_description['Market median'] = df3.groupby('Sector NAICS\nlevel 1')['Market'].median()
df_description['ebit sum'] = df3.groupby('Sector NAICS\nlevel 1')['ebit'].sum()
df_description['revenue sum'] = df3.groupby('Sector NAICS\nlevel 1')['revenue'].sum()
df_description['OPM mean'] = df_description['ebit sum'] / df_description['revenue sum']
df_description['Firms'] = df3['Sector NAICS\nlevel 1'].value_counts()
df_description.reset_index(inplace=True)
df_description
```

Out[12]:

	Sector NAICS\nlevel 1	Book	Book median	Market mean	Market median	ebit sum	revenue sum	OPM mean	Firms
0	-	5.704446e+06	5704446.000	4.865183e+06	4.865183e+06	1.267100e+05	2.732860e+05	0.463653	2
1	Accommodation and Food Services	5.139213e+05	243717.500	8.681070e+06	1.394617e+06	1.111095e+07	6.047732e+07	0.183721	69
2	Administrative and Support and Waste Managemen...	2.819477e+06	566167.000	1.385569e+07	1.938284e+06	1.909772e+07	1.218045e+08	0.156790	133
3	Agriculture, Forestry, Fishing and Hunting	3.629560e+06	1104345.000	8.046780e+06	1.264045e+06	3.251190e+06	2.197062e+07	0.147979	16
4	Arts, Entertainment, and Recreation	5.394410e+06	67242.769	1.278260e+07	2.504698e+06	4.404016e+06	3.333987e+07	0.132095	21
5	Construction	2.535521e+06	998146.500	3.857422e+06	1.745045e+06	9.855623e+06	6.709260e+07	0.146896	45
6	Educational Services	8.931767e+05	649699.000	1.302581e+06	1.524843e+06	1.861960e+05	3.567683e+06	0.052190	14
7	Finance and Insurance	5.482677e+06	1049158.000	8.412277e+06	1.264517e+06	1.716760e+08	6.407515e+08	0.267929	701
8	Health Care and Social Assistance	1.080145e+06	451385.000	3.507730e+06	1.338427e+06	-1.527342e+06	5.087670e+07	-0.030020	64
9	Information	4.213150e+06	500953.500	1.918280e+07	2.586582e+06	5.164214e+07	3.232527e+08	0.159758	261
10	Manufacturing	2.417298e+06	244642.000	1.040643e+07	5.911289e+05	2.662130e+08	1.782879e+09	0.149316	1565
11	Mining, Quarrying, and Oil and Gas Extraction	3.345942e+06	741145.000	6.783299e+06	1.042405e+06	4.406719e+07	1.190024e+08	0.370305	103
12	Other Services (except Public Administration)	5.954443e+05	431667.000	2.136156e+06	8.540956e+05	5.635980e+05	5.104295e+06	0.110416	16
13	Professional, Scientific, and Technical Services	3.566048e+06	279188.000	1.723014e+07	9.419611e+05	3.065421e+07	1.645780e+08	0.186260	145
14	Real Estate and Rental and Leasing	1.816059e+06	634398.000	3.625235e+06	8.819927e+05	7.586160e+06	5.155032e+07	0.147160	68

EN ESTA TABLA, PODEMOS VER LA MEDIA Y MEDIANA DE BOOK Y MARKET VALUE POR INDUSTRTA, EL OPM MEAN Y LA CANTIDAD DE EMPRESAS EN CADA INDUSTRIA

In [13]:
```
px.box(df3, x = 'Sector NAICS\nlevel 1',  y = "Market")
```



AQUI PODEMOS APRECIAR QUE EN EN LOS SECTORES DE MANUFACTURING, RETAIL TRADE, PROFESSIONAL SERVICES E INFORMATION, TENEMOS EMPRESAS QUE SOBRESALEN POR MUCHO DE LA MEDIA DE SUS INDUSTRIAS EN CUANTO AL MARKET VALUE, LO CUAL NOS GENERA ESTOS VALORES TAN DESVARIADOS EN EL GRAFICO

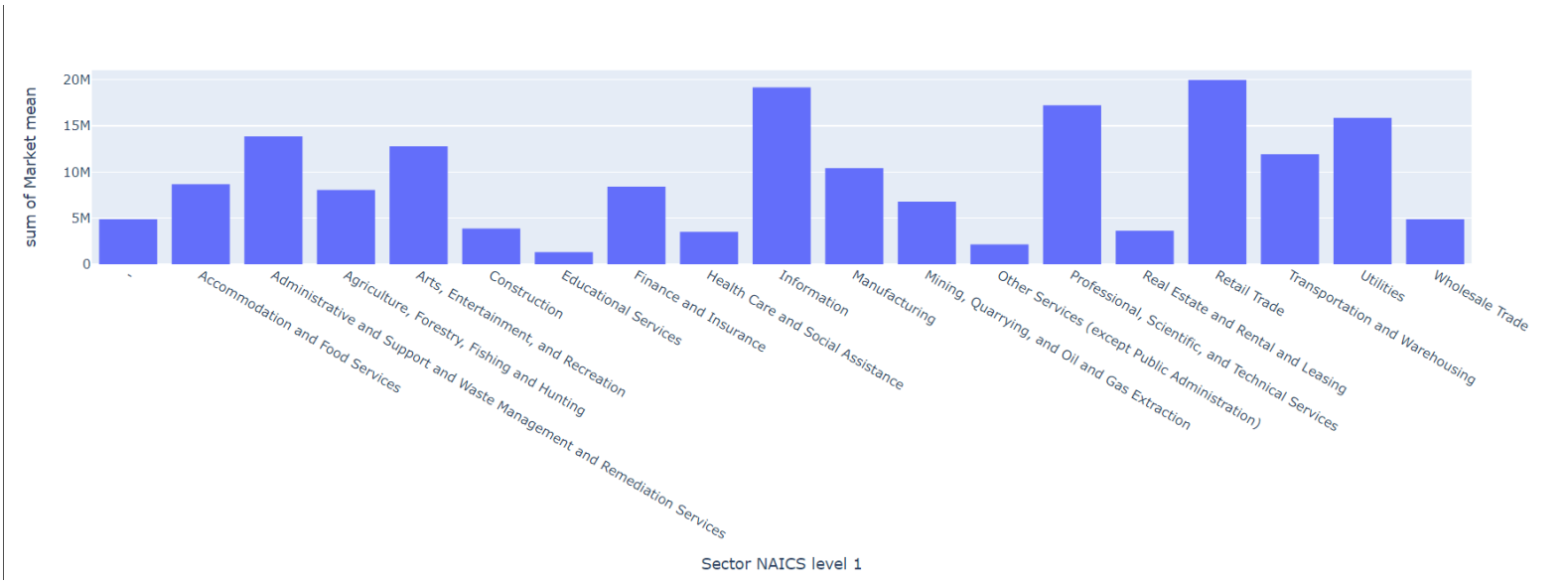
IGUALMENTE, PODEMOS VER QUE EL 75% DE LAS EMPRESAS (TODO LO QUE SE ENCUENTRA POR DEBAJO DE LA SEGUNDA RALLA DEL BOX) EN LA MAYORIA DE INDUSTRIAS NO ESTA NI CERCA DE LOS .5 BILLONES, CUANDO TENEMOS ALGUNAS EMPRESAS QUE SOBREPASAN ESTE VALOR

In [14]:
```
df3['Market'].describe()
```

Out[14]:
```
count    3.548000e+03
mean     1.095980e+07
std      6.632062e+07
min      3.490000e+01
25%      1.907523e+05
50%      1.105076e+06
75%      4.607257e+06
max      2.212838e+09
Name: Market, dtype: float64
```

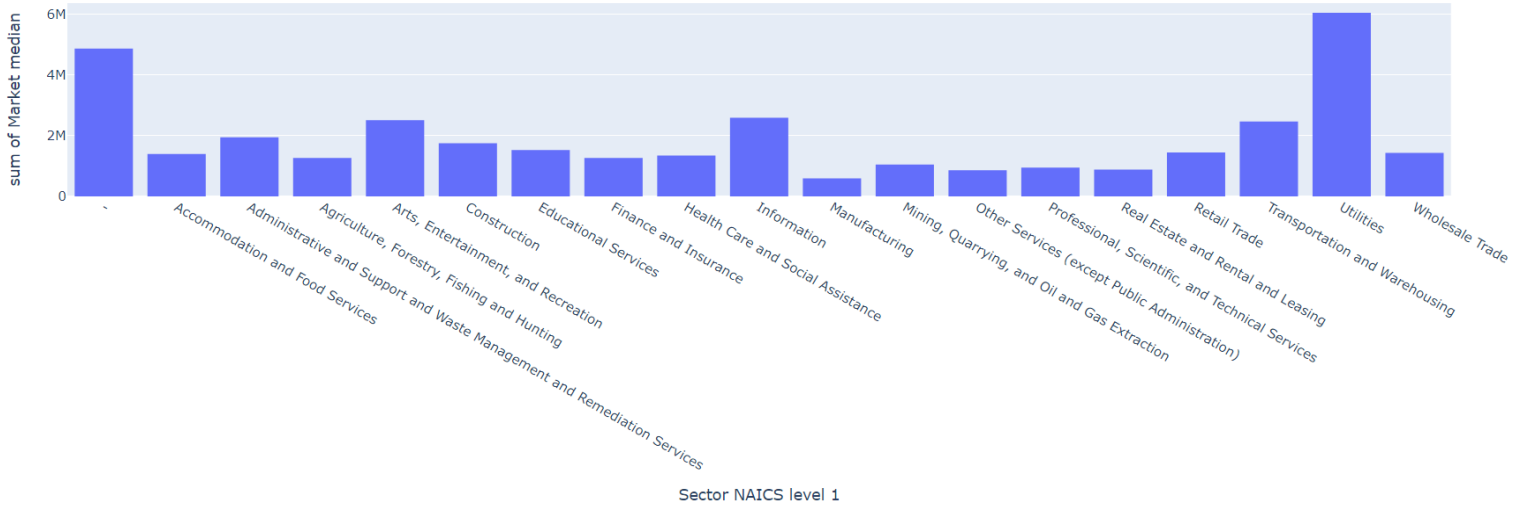
EN ESTOS DATOS, PODEMOS VER QUE LA MEDIA DEL MARKET VALUE ES MUCHO MAYOR QUE EL 75% DE LAS EMPRESAS, LO CUAL NOS QUIERE DECIR QUE TENEMOS UNAS CUANTAS EMPRESAS CON VALORES ALTISIMOS DE MARKET VALUE, LO CUAL HACE QUE ESTA MEDIDA NO SEA REPRESENTATIVA PARA LAS EMPRESAS, EN LUGAR DE ESTA SE DEBERIA TOMAR LA MEDIANA, YA QUE MA MAYORIA DE LAS EMPRESAS TIENEN UN VALOR MAS BAJO QUE LA MEDIA

In [15]:
```
px.histogram(df_description, x = 'Sector NAICS\nlevel 1', y = 'Market mean')
```



AQUI PODEMOS VER COMO SE DISTRIBUYE EL MEAN MARKET VALUE POR INDUSTRIA, POR LO QUE VEMOS QUE EXISTEN INDUSTRIAS COMO INFORMATICA, RETAIL TRADE Y PROFESSIONAL, SCIENTIFIC AND TECNICAL SERVICES, LOS CUALES TIENEN UNA MEDIA MUY ALTA EN MARKET VALUE, A DIFERENCIA DE OTROS SECTORES COMO EDUCATIONAL SERVICES, REAL ESTATE AND RENTAL, Y HEALTH CARE ANS SOCIAL ASSISTANCE, LO CUAL ES CURIOSO, DEBIDO A QUE ESTAS INDUSTRIAS TRATAN MAS SOBRE EL CUIDADO DE LA SALUD

In [16]:
```
px.histogram(df_description, x = 'Sector NAICS\nlevel 1', y = 'Market median')
```

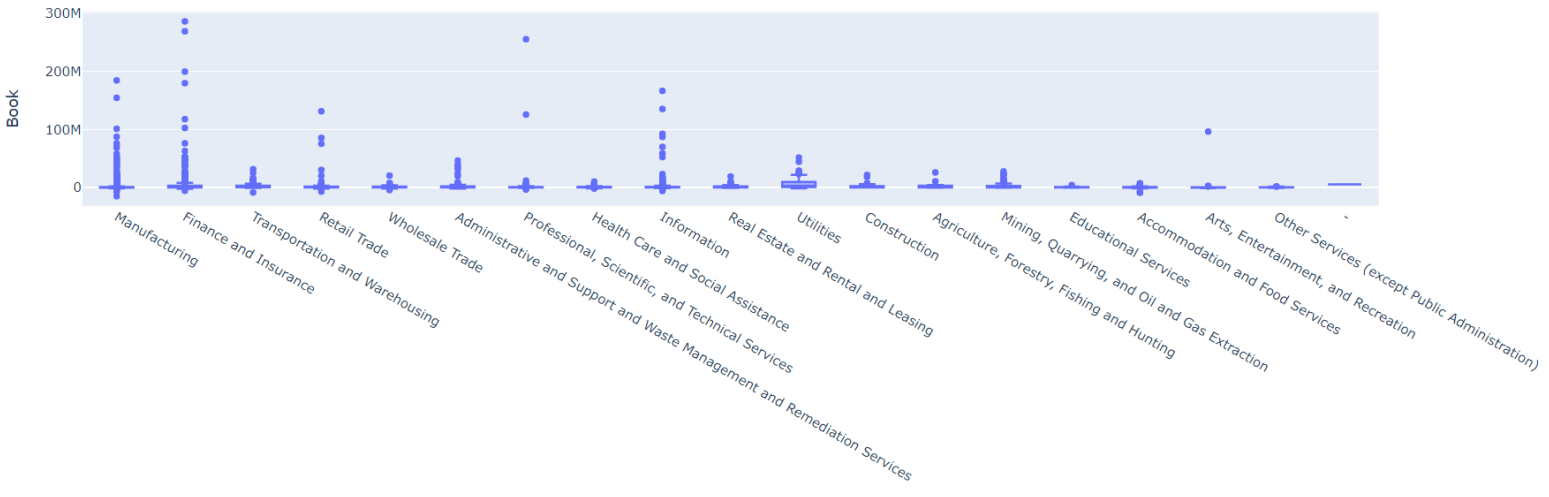


In [17]:
```
market_mean = df_description['Market median'].mean()
market_mean
```

Out[17]:
```
1900659.9883757897
```

EN ESTE OTRO GRAFICO, PODEMOS VER ESTA MISMA VARIABLE, PERO ESTA VEZ TOMANDO LA MEDIANA, Y PODEMOS VER QUE LA DISTRIBUCION ES MUY DIFERENTE, LOS VALORES MAS ALTOS ENCONTRADOS SON DE MENOS DE 6M, Y LAS INDUSTRIAS DE MAYOR TAMAÑO SON -, INFORMATION Y UTILITIES, IGUAL, PODEMOS VER QUE LA EMPRESA TIPICA DE EU TIENE UN MARKET VALUE DE 1,900,659,988 DOLARES

In [18]:
```
px.box(df3, x = 'Sector NAICS\nlevel 1',  y = "Book")
```



AQUI PODEMOS APRECIAR QUE EN EN LOS SECTORES DE MANUFACTURING, FINANCE AND INSURANCE, PROFESSIONAL SERVICES, INFORMATION ENTRE OTRAS, TENEMOS EMPRESAS QUE SOBRESALEN POR MUCHO DE LA MEDIA DE SUS INDUSTRIAS EN CUANTO AL BOOK VALUE, LO CUAL NOS GENERA ESTOS VALORES TAN DESVARIADOS EN EL GRAFICO

IGUALMENTE, PODEMOS VER QUE EL 75% DE LAS EMPRESAS (TODO LO QUE SE ENCUENTRA POR DEBAJO DE LA SEGUNDA RALLA DEL BOX) EN LA MAYORIA DE INDUSTRIAS NO ESTA NI CERCA DE LOS 100 MILLONES, CUANDO TENEMOS ALGUNAS EMPRESAS QUE SOBREPASAN ESTE VALOR

In [19]:
```
df3['Book'].describe()
```

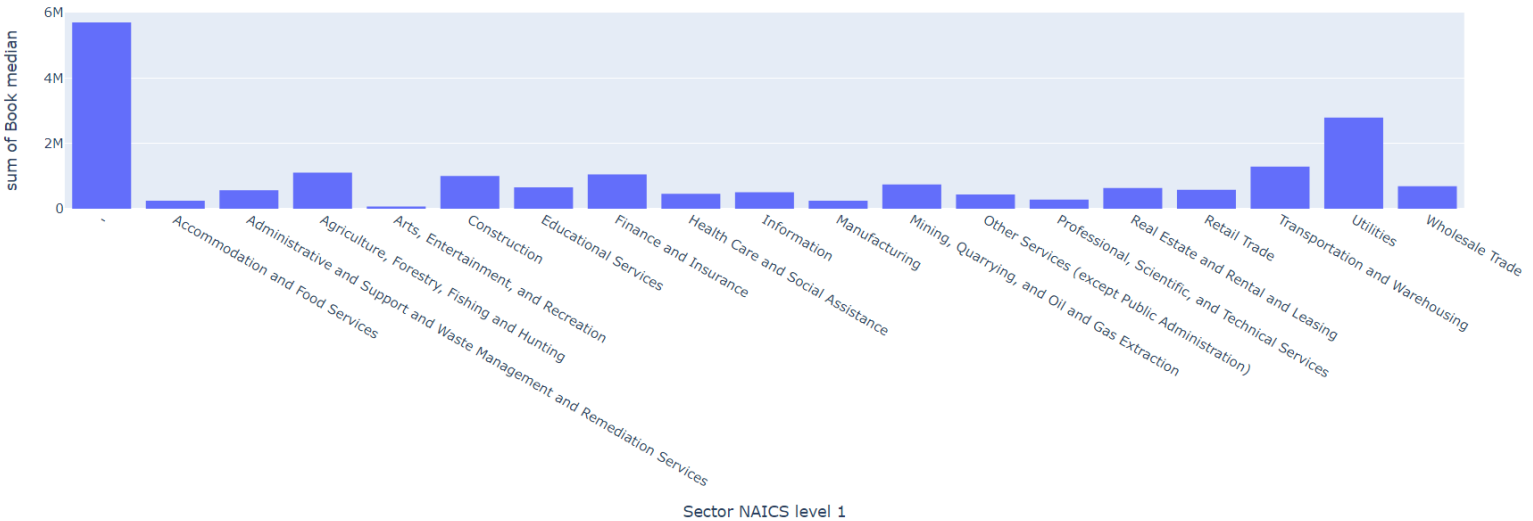
Out[19]:
```
count    3.362000e+03
mean     3.331749e+06
std      1.386098e+07
min     -1.479100e+07
25%      1.060878e+05
50%      4.577370e+05
75%      1.792766e+06
max      2.861430e+08
Name: Book, dtype: float64
```

EN ESTOS DATOS, PODEMOS VER QUE LA MEDIA DEL BOOK VALUE ES MUCHO MAYOR QUE EL 75% DE LAS EMPRESAS, LO CUAL NOS QUIERE DECIR QUE TENEMOS UNAS CUANTAS EMPRESAS CON VALORES ALTISIMOS DE BOOK VALUE, LO CUAL HACE QUE ESTA MEDIDA NO SEA REPRESENTATIVA PARA LAS EMPRESAS, EN LUGAR DE ESTA SE DEBERIA TOMAR LA MEDIANA, YA QUE MA MAYORIA DE LAS EMPRESAS TIENEN UN VALOR MAS BAJO QUE LA MEDIA

In [20]:
```
px.histogram(df_description, x = 'Sector NAICS\nlevel 1', y = 'Book')
```

AQUI PODEMOS VER COMO SE DISTRIBUYE EL BOOK MARKET VALUE POR INDUSTRIA, POR LO QUE VEMOS LA MAYORIA DE ESTAS EMOPRESAS TIENEN UNA MEDIA CERCANA A 4M +- 2M, AUNQUE EXISTEN ALGUNAS NDISTRIAS COMO EDUCATIONAL SERVICES, HEALTH CARE AND SOCIAL ASSISTANCE LOS CUALES SON MENORES O ACCOMMODATION AND FOOD SERVICES

In [21]:
```
px.histogram(df_description, x = 'Sector NAICS\nlevel 1', y = 'Book median')
```



In [22]:
```
book_mean = df_description['Book median'].mean()
book_mean
```

Out[22]:
```
1000405.1720526316
```

EN ESTE OTRO GRAFICO, PODEMOS VER ESTA MISMA VARIABLE, PERO ESTA VEZ TOMANDO LA MEDIANA, Y PODEMOS VER QUE LA DISTRIBUCION ES DIFERENTE, LOS VALORES MAS ALTOS ENCONTRADOS SON DE MENOS DE 6M, Y LAS INDUSTRIAS DE MAYOR TAMAÑO SON - Y UTILITIES, IGUAL, PODEMOS VER QUE LA EMPRESA TIPICA DE EU TIENE UN BOOK VALUE DE 1,000,405,172 DOLARES

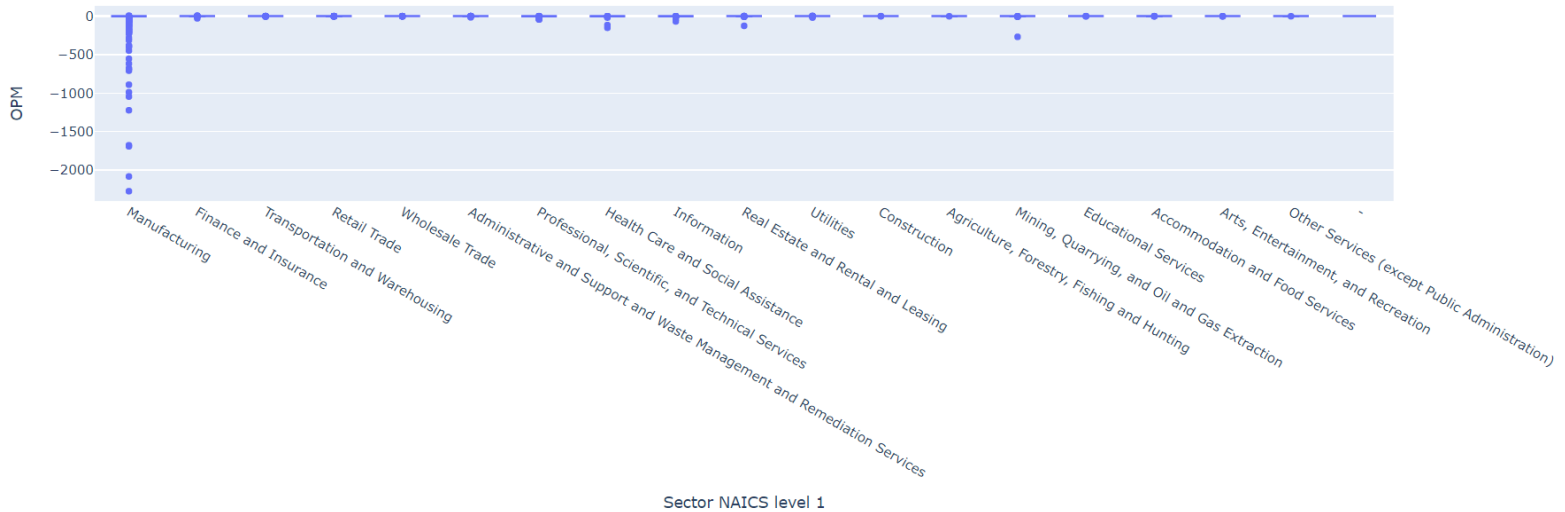
For each industry (and for all industries), what can you say about profit margin of firms? show a) descriptive statistics of profit margin and b) plot(s) to illustrate how profit margin changes across industries.

In [23]:
```
df_description
```

Out[23]:

	Sector NAICS\nlevel 1	Book	Book median	Market mean	Market median	ebit sum	revenue sum	OPM mean	Firms
0	-	5.704446e+06	5704446.000	4.865183e+06	4.865183e+06	1.267100e+05	2.732860e+05	0.463653	2
1	Accommodation and Food Services	5.139213e+05	243717.500	8.681070e+06	1.394617e+06	1.111095e+07	6.047732e+07	0.183721	69
2	Administrative and Support and Waste Managemen...	2.819477e+06	566167.000	1.385569e+07	1.938284e+06	1.909772e+07	1.218045e+08	0.156790	133
3	Agriculture, Forestry, Fishing and Hunting	3.629560e+06	1104345.000	8.046780e+06	1.264045e+06	3.251190e+06	2.197062e+07	0.147979	16
4	Arts, Entertainment, and Recreation	5.394410e+06	67242.769	1.278260e+07	2.504698e+06	4.404016e+06	3.333987e+07	0.132095	21
5	Construction	2.535521e+06	998146.500	3.857422e+06	1.745045e+06	9.855623e+06	6.709260e+07	0.146896	45
6	Educational Services	8.931767e+05	649699.000	1.302581e+06	1.524843e+06	1.861960e+05	3.567683e+06	0.052190	14
7	Finance and Insurance	5.482677e+06	1049158.000	8.412277e+06	1.264517e+06	1.716760e+08	6.407515e+08	0.267929	701
8	Health Care and Social Assistance	1.080145e+06	451385.000	3.507730e+06	1.338427e+06	-1.527342e+06	5.087670e+07	-0.030020	64
9	Information	4.213150e+06	500953.500	1.918280e+07	2.586582e+06	5.164214e+07	3.232527e+08	0.159758	261
10	Manufacturing	2.417298e+06	244642.000	1.040643e+07	5.911289e+05	2.662130e+08	1.782879e+09	0.149316	1565
11	Mining, Quarrying, and Oil and Gas Extraction	3.345942e+06	741145.000	6.783299e+06	1.042405e+06	4.406719e+07	1.190024e+08	0.370305	103
12	Other Services (except Public Administration)	5.954443e+05	431667.000	2.136156e+06	8.540956e+05	5.635980e+05	5.104295e+06	0.110416	16
13	Professional, Scientific, and Technical Services	3.566048e+06	279188.000	1.723014e+07	9.419611e+05	3.065421e+07	1.645780e+08	0.186260	145
14	Real Estate and Rental and Leasing	1.816059e+06	634398.000	3.625235e+06	8.819927e+05	7.586160e+06	5.155032e+07	0.147160	68
15	Retail Trade	3.379848e+06	577426.500	1.998810e+07	1.444971e+06	4.110833e+07	7.570414e+08	0.054301	152
16	Transportation and Warehousing	3.536723e+06	1288121.500	1.191248e+07	2.464494e+06	2.030547e+07	1.816330e+08	0.111794	69
17	Utilities	7.148295e+06	2791950.000	1.588301e+07	6.040649e+06	1.544359e+07	1.167458e+08	0.132284	77
18	Wholesale Trade	1.529928e+06	683900.000	4.871519e+06	1.424602e+06	1.045879e+07	3.338191e+08	0.031331	79

In [24]:
```
px.box(df3, x = "Sector NAICS\nlevel 1", y = 'OPM')
```



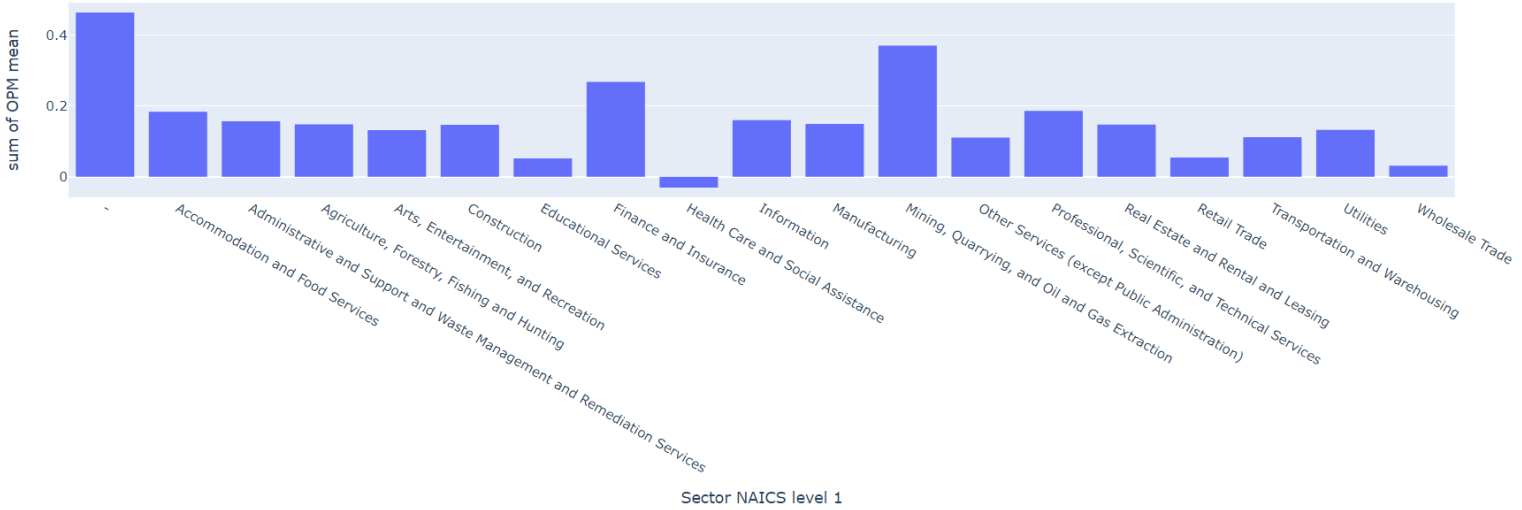
EN ESTE GRAFICO, SE PUEDE VER EL OPM POR INDUSTRIA Y SUS VALORES SEGUN LAS EMPRESAS DE CADA INDUSTRIA, PODEMOS VER QUE EL 75% DE LAS EMPRESAS POR INDUSTRIA ESTA MUY CERCA A 0, REALMENTE HAY POCAS EMPRESAS CON VALORES ATIPICOS, PERO EXISTEN Y TIENEN UN OPM MUY BAJO, TENDIENDO A PASAR MAS EN LA INDUSTRIA DE MANUFACTURING

In [25]:
```
df3['OPM'].describe()
```

Out[25]:
```
count    3354.000000
mean             NaN
std              NaN
min             -inf
25%        -0.258106
50%         0.068118
75%         0.211226
max              inf
Name: OPM, dtype: float64
```

EN ESTE CASO, DEBIDO A QUE PARA OBTENER EL OPM DEBEMOS REALIZAR UNA DIVICION, CUANDO EL DENOMINADOR, QUE EN ESTE CASO SON LAS SALES, ES 0, OBTENEMOS INDEFINIDO , POR LO QUE NO TENEMOS MINIMO O MAXIMO DEBIDO A LA FALTA DE DATOS, POR LO MISMO NUESTRO PROMEDIO NO ESTA DEFINIDO, PERO TENEMOS EL VALOR DEL 50% (LA MEDIANA) QUE ES DE .068 Y ES UN VALOR QUE PODEMOS TOMAR

In [26]:
```
px.histogram(df_description, x = 'Sector NAICS\nlevel 1', y = 'OPM mean')
```



EN ESTE GRAFICO, TENEMOS LA MEDIA SEGUN LA INDUSTRIA, COMO YA MENCIONE, HAY CASOS DONDE NO PODEMOS CONOCER LA MEDIA DEBIDO A LA FALTA DE DATOS, PERO PODEMOS VER QUE LA MAYORIA TIENE UN OPM MAYOR A 0 A EXCEPCION DE ADMINISTRATIVE AND SUPPORT ASN WASTE MANAGEMENT, EL CUAL TIENE UN OPM NEGATIVO

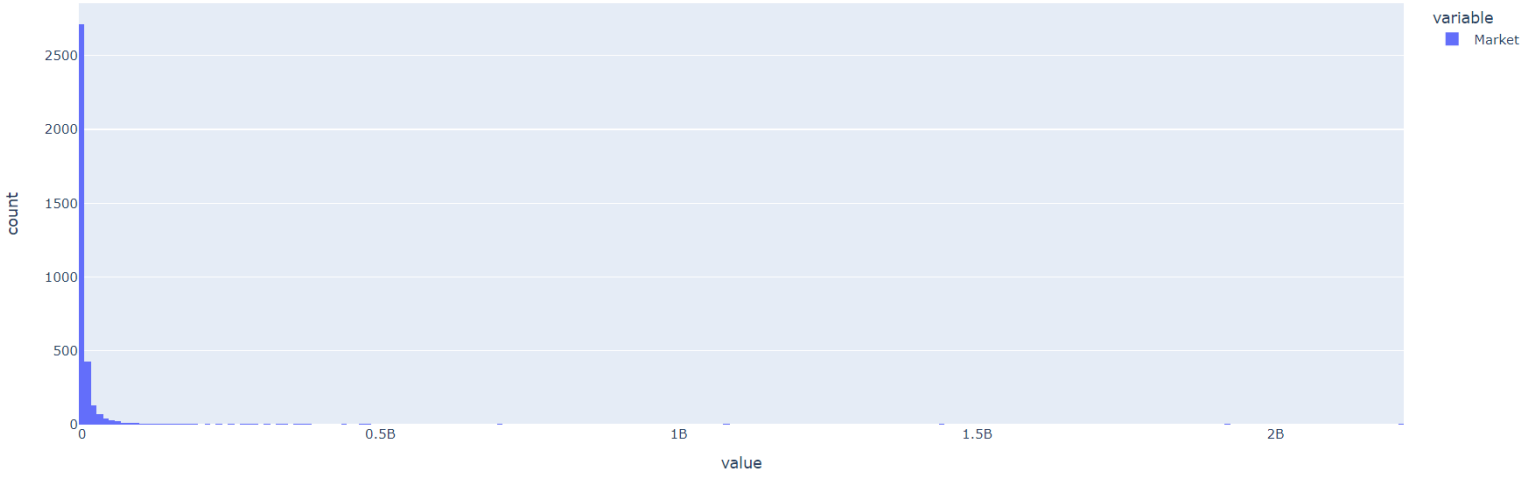
Which are the biggest 10 US firms in terms of market value and how far they are from the typical size of a US firm?

In [27]:
```
df3.sort_values('Market', ascending=False).head(10)
```

Out[27]:

	firm	q	revenue	cogs	sgae	otheropexp	extraincome	finexp	incometax	totalassets	...	Class	Sector\nAICS\nlevel 1	Exchange / Src	Sector\nEconomati
809	AAPL	2022q2	82959000.0	47074000.0	12809000.0	0.0	-10000.0	0.0	3624000.0	336309000.0	...	Com	Manufacturing	NASDAQ	Electric Electr
191175	MSFT	2022q2	51865000.0	16429000.0	14902000.0	0.0	-47000.0	0.0	3747000.0	364840000.0	...	Com	Information	NASDAQ	Software & Da
125851	GOOGL	2022q2	69685000.0	30104000.0	20128000.0	0.0	-439000.0	0.0	3012000.0	355185000.0	...	Com A	Professional, Scientific, and Technical Services	NASDAQ	Oth
18173	AMZN	2022q2	121234000.0	66424000.0	51403000.0	90000.0	-5557000.0	425000.0	-637000.0	419728000.0	...	Com	Retail Trade	NASDAQ	Trac
289525	TSLA	2022q2	16934000.0	12700000.0	1628000.0	142000.0	18000.0	18000.0	205000.0	68513000.0	...	Com	Manufacturing	NASDAQ	Vehicle & Par
296815	UNH	2022q2	80332000.0	73200000.0	0.0	0.0	-129000.0	467000.0	1466000.0	230172000.0	...	Com	Finance and Insurance	NYSE	Finance and Insuran
156887	JNJ	2022q2	24020000.0	7919000.0	9929000.0	85000.0	-273000.0	-26000.0	1026000.0	177724000.0	...	Com	Manufacturing	NYSE	Chemic
182535	META	2022q2	28822000.0	5192000.0	15272000.0	0.0	-172000.0	0.0	1499000.0	169779000.0	...	Com A	Professional, Scientific, and Technical Services	NASDAQ	Oth
205565	NVDA	2022q2	8288000.0	2857000.0	2210000.0	1353000.0	-13000.0	50000.0	187000.0	45212000.0	...	Com	Manufacturing	NASDAQ	Electric Electr
300325	V	2022q2	7275000.0	0.0	3127000.0	0.0	-208000.0	111000.0	418000.0	85410000.0	...	Com A	Administrative and Support and Waste Managemen...	NYSE	Oth

In [28]:
```
px.histogram(df3['Market'])
```



In [29]:
```
print(df3['Market'].median())
```

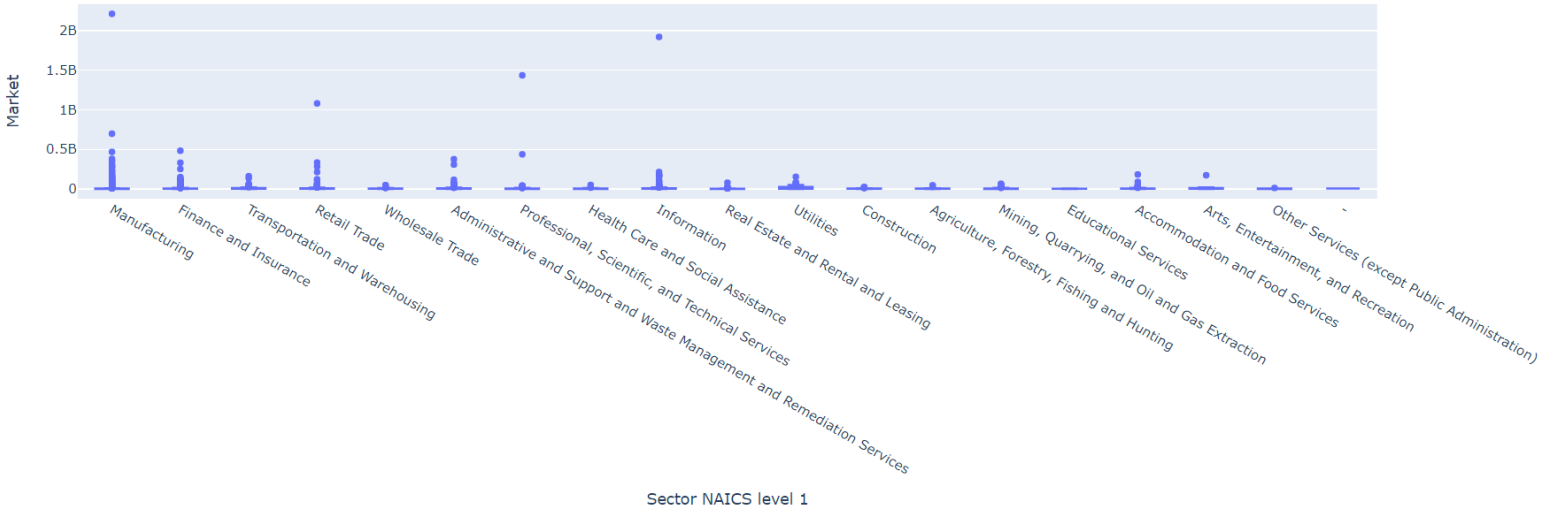
```
1105075.91083
```

PARA OBTENER LAS 10 FIRMAS CON MAYOR MARKET VALUE, PRIMERO SE FILTRO PARA UNICAMENTE TENER LOS DATOS MAS RECIENTES Y ASI CONOCER EL MARKET VALUE ACTUAL DE LAS FIRMAS. DESPUES SE ORDENAN LOS DATOS DE MAYOR A MENOR Y OBTUVIMOS LAS 10 FIRMAS DE US CON MAYOR MARKET VALUE, QUE SON:

- AAPL
- MSFT
- GOOGL
- AMZN
- TSLA
- UNH
- JNJ
- META
- NVDA
- V

EN CUANTO A QUE TAN DESVIADOS ESTAN DE EL VALOR TIPICO, AL REALIZAR UN HISTOGRAMA PODEMOS VER QUE EL VALOR ESTA MUY SESGADO A LA IZQUIERDA, POR LO QUE EN LUGAR DE USAR LA MEDIA USAREMOS LA MEDIANA QUE ES 1105075.91083

In [30]:
```
px.box(df3, x = "Sector NAICS\nlevel 1", y = 'Market')
```



FINALMENTE PARA CONOCER COMO SE COMPORTAN POR CADA SECTOR, PODEMOS VER EL SIGUIENTE GRAFICO, DONDE VEMOS QUE TENEMOS SECTORES QUE TIENEN VALORES DE MARKET MUY ALTOS, PERO REALMENTE ESO SOLO SON PICOS DENTRO DE LOS SECTORES, CUANDO LA MAYORIA SE ENCUENTRA MAS CERCANA A 0

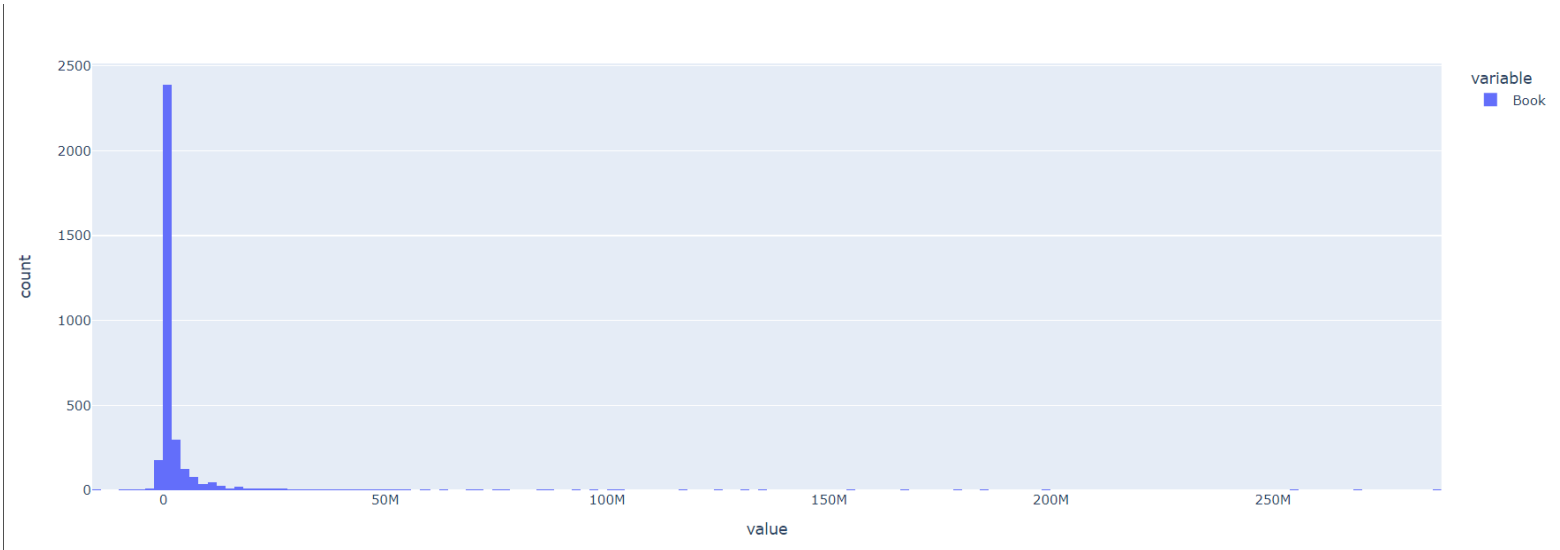
Which are the biggest 10 US firms in terms of book value and how far they are from the typical size of a US firm?

In [31]:
```
df3.sort_values('Book', ascending=False).head(10)
```

Out[31]:

	firm	q	revenue	cogs	sgae	otheropexp	extraincome	finexp	incometax	totalassets	...	Class	Sector\nNAICS\nlevel 1	Exchange / Src	Sector\nEconom
157427	JPM	2022q2	18646000.0	3518000.0	0.0	0.0	-4263000.0	0.0	2216000.0	3.841314e+09	...	Com	Finance and Insurance	NYSE	Finance and Insur
34013	BAC	2022q2	14975000.0	2531000.0	0.0	0.0	-5552000.0	0.0	645000.0	3.111606e+09	...	Com	Finance and Insurance	NYSE	Finance and Insur
125851	GOOGL	2022q2	69685000.0	30104000.0	20128000.0	0.0	-439000.0	0.0	3012000.0	3.551850e+08	...	Com A	Professional, Scientific, and Technical Services	NASDAQ	C
49680	C	2022q2	15630000.0	3666000.0	0.0	0.0	-6235000.0	0.0	1182000.0	2.380904e+09	...	Com	Finance and Insurance	NYSE	Finance and Insur
319405	XOM	2022q2	111265000.0	76299000.0	6981000.0	7154000.0	3572000.0	194000.0	6359000.0	3.677740e+08	...	Com	Manufacturing	NYSE	Oil &
312205	WFC	2022q2	11556000.0	1358000.0	0.0	0.0	-6466000.0	0.0	613000.0	1.881142e+09	...	Com	Finance and Insurance	NYSE	Finance and Insur
191175	MSFT	2022q2	51865000.0	16429000.0	14902000.0	0.0	-47000.0	0.0	3747000.0	3.648400e+08	...	Com	Information	NASDAQ	Software &
78332	CVX	2022q2	68762000.0	46321000.0	4563000.0	1759000.0	-80000.0	129000.0	4288000.0	2.579360e+08	...	Com	Manufacturing	NYSE	Oil &
277650	T	2022q2	29643000.0	12341000.0	11715000.0	631000.0	2212000.0	1502000.0	1509000.0	4.264330e+08	...	Com	Information	NYSE	Telecommunic
18173	AMZN	2022q2	121234000.0	66424000.0	51403000.0	90000.0	-5557000.0	425000.0	-637000.0	4.197280e+08	...	Com	Retail Trade	NASDAQ	T

In [32]:
```
px.histogram(df3['Book'])
```



In [33]:
```
print(df3['Book'].median())
```

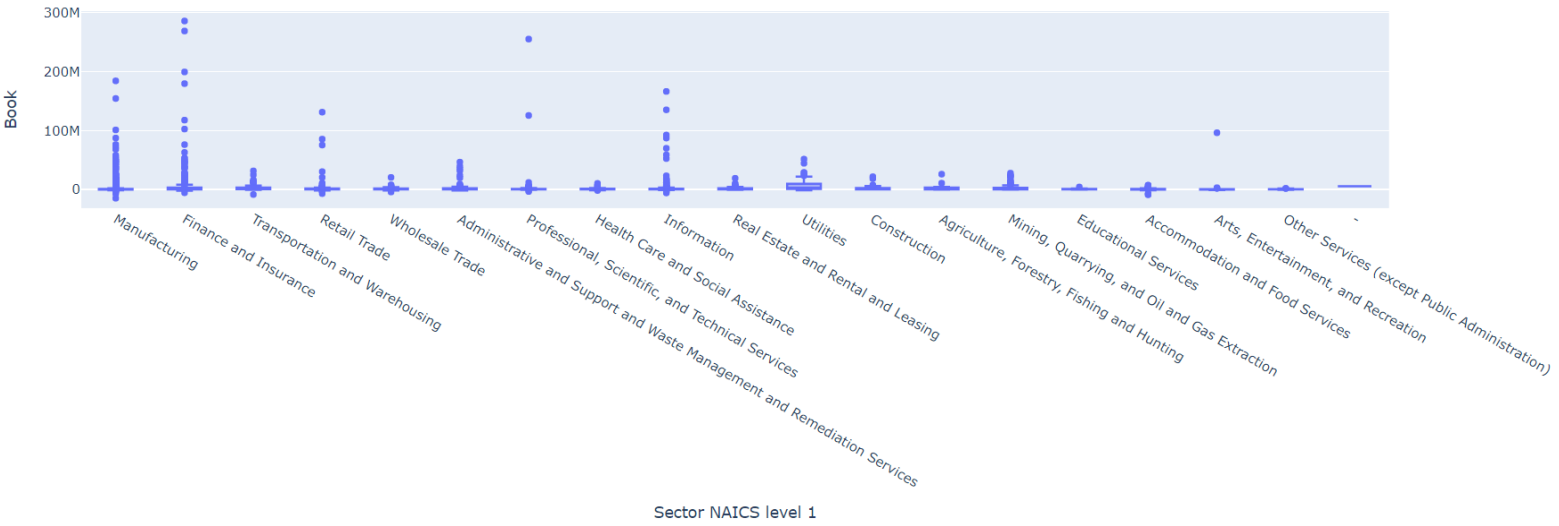
```
457737.0
```

PARA OBTENER LAS 10 FIRMAS CON MAYOR BOOK VALUE, PRIMERO SE FILTRO PARA UNICAMENTE TENER LOS DATOS MAS RECIENTES Y ASI CONOCER EL BOOK VALUE ACTUAL DE LAS FIRMAS. DESPUES SE ORDENAN LOS DATOS DE MAYOR A MENOR Y OBTUVIMOS LAS 10 FIRMAS DE US CON MAYOR BOOK VALUE, QUE SON:

- JPM
- BAC
- GOOGL
- C
- XOM
- WFC
- MSFT
- CVX
- T
- AMZN

EN CUANTO A QUE TAN DESVIADOS ESTAN DE EL VALOR TIPICO, AL REALIZAR UN HISTOGRAMA PODEMOS VER QUE EL VALOR ESTA MUY SESGADO A LA IZQUIERDA, POR LO QUE EN LUGAR DE USAR LA MEDIA USAREMOS LA MEDIANA QUE ES 457737.0

In [34]:
```
px.box(df3, x = "Sector NAICS\nlevel 1", y = 'Book')
```



FINALMENTE PARA CONOCER COMO SE COMPORTAN POR CADA SECTOR, PODEMOS VER EL SIGUIENTE GRAFICO, DONDE VEMOS QUE TENEMOS SECTORES QUE TIENEN VALORES DE BOOK MUY ALTOS, PERO REALMENTE ESO SOLO SON PICOS DENTRO DE LOS SECTORES, CUANDO LA MAYORIA SE ENCUENTRA MAS CERCANA A 0

2.2.1.2

How can you measure firm profitability that can be used to compare performance among firms of different sizes? Select and justify at least 3 measures and show descriptive statistics

PARA CONOCER LA FIRMA CON MAYOR PROFITABILITY, PODEMOS TOMAR LOS EARNING PEAR SHARE DEFLATED BY PRICE, EL CUAL NOS DICE LAS GANANCIAS DE UNA EMPRESESA SEGUN LO QUE VENDE, OPM, QUE HABLA SOBRE EL PORCENTAJE DE GANANCIA DE SUS VENTAS Y BOOK TO MARKET RATIO, QUE ES UNA MEDIDA DE COMPARACION ENTRE EL BOOK VALUE Y EL MARKET VALUE

Calculate and explain earnings per share deflated by price.

FORMULAS PARA OBTENER EL EPSP

- ebit = revenue - cogs - sgae - otheropexp
- Net Income = ebit - incometax - finexp
- Earnings pear share = EPS = netIncome/#sharesoutstanding
- EPSP = EPS / stockPrice = EPS deflated by price

EL EPS ES EL INDICADOR DE LA RENTABILIDAD DE UNA EMPRESA, ESTO YA QUE SE OBTIENE CON LAS GANANCIAS NETA ENTRE LAS ACCIONES EN CIRCULACION, CONOCIENDO EL VALOR REAL DE CADA ACCION

2.2.2

In [35]:
```
df_sort = df1.copy()
```

In [36]:
```
df_sort_merge = df_sort.merge(df2, left_on='firm', right_on='Ticker')
```

In [37]:
```
df_sort_merge['Market'] = df_sort_merge['originalprice'] * df_sort_merge['sharesoutstanding']
```

In [38]:
```
df_sort_merge.replace([np.inf, -np.inf], np.nan, inplace=True)
```

In [39]:
```
df_sort_merge.dropna(subset=["Market"], how="all", inplace=True)
```

CLASIFICAMOS LAS EMPRESAS SEGUN SU TAMAÑO SEGUN EL TRIMESTRE EN EL QUE SE ENCONTRABAN

In [40]:
```
def dense_inclusive_pct(x):
    # I subtract one to handle the inclusive bit
    r = x.rank(method='dense') - 1
    return r / r.max() * 100

df_sort_merge["pct"]=df_sort_merge.groupby('q')['Market'].apply(dense_inclusive_pct).astype(int)

#df_analysis[["q","Market","pct"]].sort_values("q")

df_sort_merge["isSmall"] = df_sort_merge.pct <= 33
df_sort_merge["isSmall"] = df_sort_merge["isSmall"].astype(int)

df_sort_merge["isMedium"] = (df_sort_merge.pct <= 66) & (df_sort_merge.pct > 33)
df_sort_merge["isMedium"] = df_sort_merge["isMedium"].astype(int)
```

You have to select a group of firms according to their general industry classification:

Service industries

In [41]:
```
df2_services = df_sort_merge.copy()
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Manufacturing')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Finance and Insurance')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Information')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Retail Trade')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Mining, Quarrying, and Oil and Gas Extraction')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Wholesale Trade')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Utilities')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Transportation and Warehousing')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Real Estate and Rental and Leasing')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Health Care and Social Assistance')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Construction')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Arts, Entertainment, and Recreation')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == '-')].index)
df2_services = df2_services.drop(df2_services[(df2_services['Sector NAICS\nlevel 1'] == 'Agriculture, Forestry, Fishing and Hunting')].index)
```

REALIZAMOS LIMPIEZA DE DATOS PARA BORRAR TODAS LOS NAN y 0 DE NUESTRO DATASET

In [42]:
```
df_clean = df2_services.copy()
```

AL VER QUE TENEMOS VALORES EN REVENUE = 0, DECIDO CAMBIAR ESTOS POR NAN, PARA ASI AL CALULAR EL OPM NO TENGAMOS PROBLEMAS DEBIDO A QUE PARA OBTENER ESTE, SE REALIZA UNA DIVICION ENTRE REVENUE Y SI DIVIDIERAMOS ENTRE 0 NOS DARIAN VALORES INDEFINIDOS, LO CUAL CAUSARA RUIDO EN NUESTRO ANALISIS

In [43]:
```
df_clean['revenue'] = df_clean['revenue'].replace([0], [np.nan])
```

DROPEAMOS COLUMNAS QUE NO USAREMOS EN EL ANALISIS

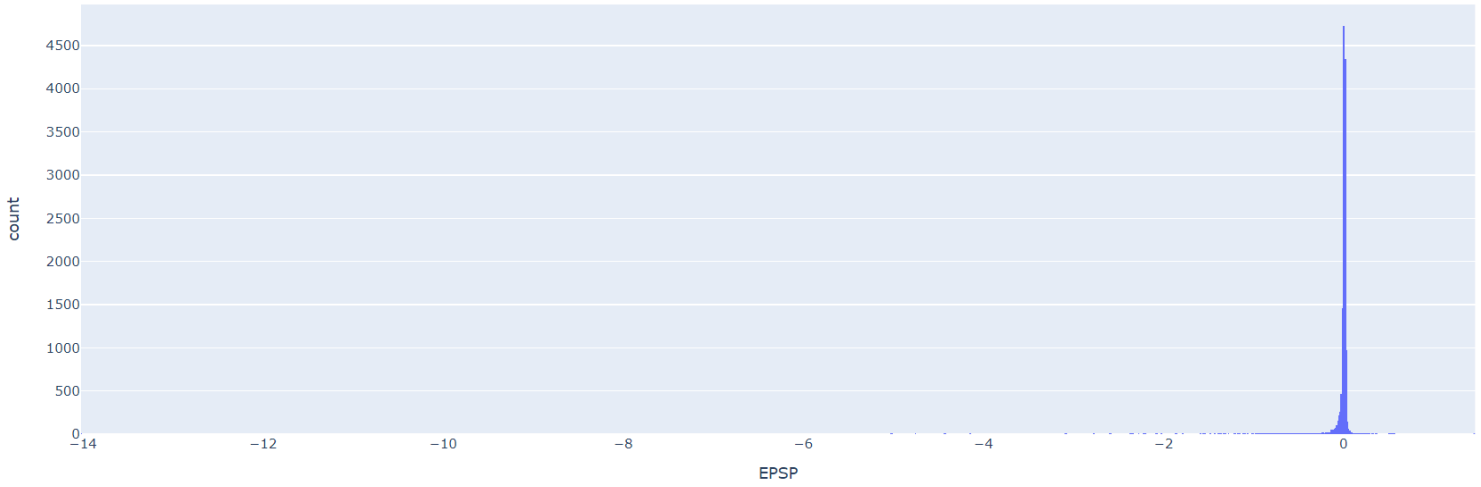
In [44]:
```
df_clean = df_clean.drop(
    [
        'extraincome',
        'shortdebt',
        'longdebt',
        'stockholderequity',
        'Ticker',
        'Name',
        'Class',
        'Exchange / Src',
        'Sector\nEconomatica',
        'Sector NAICS\nlast available',
        'partind'
      ],
       axis=1)
```

Using your subset of firms that belong to your industry, which factors (variables) might be related to annual stock return one quarter in the future? Select at least 3 factors and briefly explain why you think might be related to stock returns.

USAREMOS LAS VARIABLES DE ESPS, OPM Y BOOK TO MARKET RATIO PARA NUESTRO MODELO, YA QUE EN ESTAS PODEMOS CONOCER TANTO EL VALOR DE UNA EMPRESA SEGUN SUS VENTAS, COMO LA PROPORCION ENTRE BOOK Y MARKET VALUE DE LAS EMPRESAS QUE SON PARTE DE "SERVICE INDUSTRIES".

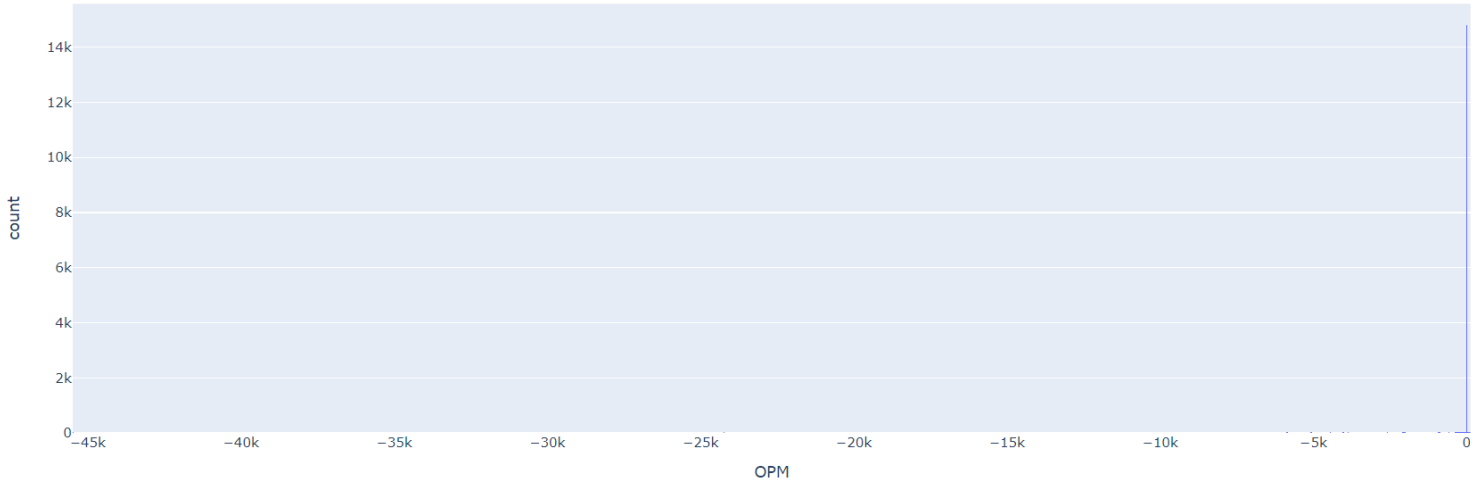
In [45]:
```
df_clean['Book'] = df_clean['totalassets'] - df_clean['totalliabilities']
df_clean['Market'] = df_clean['originalprice'] * df_clean['sharesoutstanding']


df_clean['EBIT'] = df_clean['revenue'] - df_clean['cogs'] - df_clean['sgae'] - df_clean['otheropexp']
df_clean['NetIncome'] = df_clean['EBIT'] - df_clean['incometax'] - df_clean['finexp']
df_clean['EPS'] = df_clean['NetIncome'] / df_clean['sharesoutstanding']


df_clean['EPSP'] = df_clean['EPS'] / df_clean['originalprice']
df_clean['OPM'] = df_clean['EBIT'] / df_clean['revenue']
df_clean['Book_to_Market_ratio'] = df_clean['Book'] / df_clean['Market']
```

In [46]:
```
px.histogram(df_clean, x = 'EPSP')
```



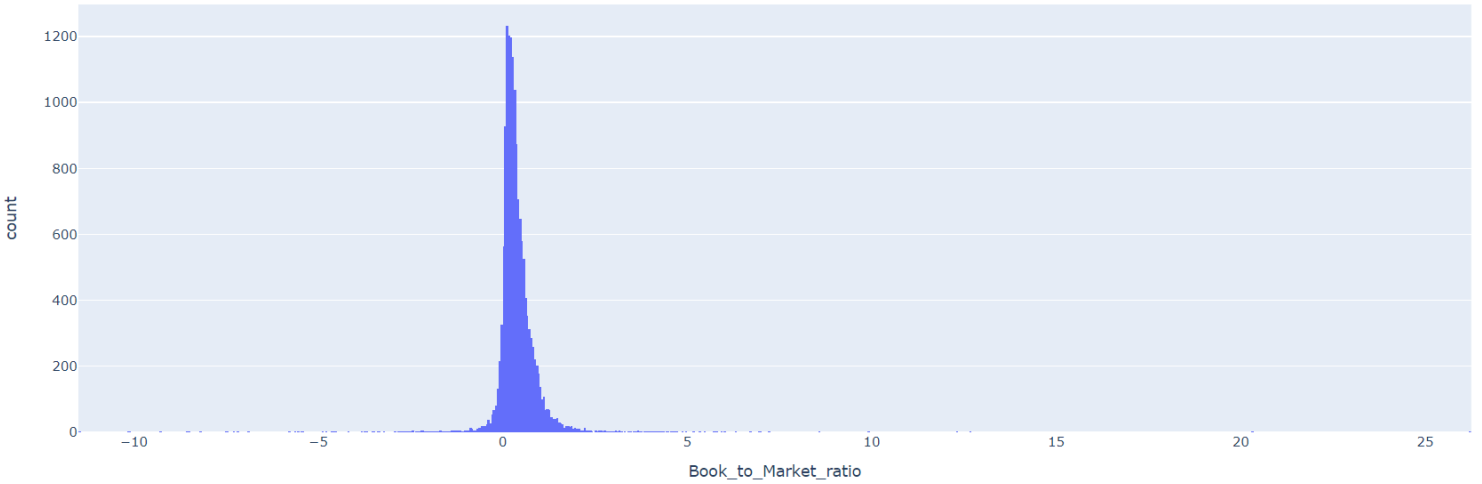
CON ESTA GRAFICA, PODEMOS VER COMO TENEMOS UN HISTOGRAMA INCLINADO HACIA LA DERECHA, YA QUE TENEMOS MUCHAS EMPRESAS CON UN VALOR DE EPSP MUY CERCANO A 0, PERO TAMBIEN EXISTEN ALGUNAS (AUNQUE MUY POCAS) CON VALORES NEGATIVOS QUE LLEGAHN HASTA EL -14. DE ESTE GRAFICO PODEMOS INFERIR QUE UN DATO REPRESENTATIVO PARA ESTA VARIABLE SERIA LA MEDIANA EN LUGAR DE LA MEDIA, DEBIDO A ESTA INCLINACION

In [47]:
```
px.histogram(df_clean, x = 'OPM')
```



AL IGUAL QUE EN LA GRAFICA ANTERIOR, PODEMOS VER UNA INCLINACION DE LOS DATOS HACIA LA DERECHA, YA QUE NUEVAMENTE LOS VALORES SE ASEMEJAN EN SU MAYORIA A 0, CUANDO EXISTEN ALGUNOS CASOS DONDE EL VALOR LLEGA A SER DE HASTA -45000 EN LA VARIABLE DE OPERATING PROFIT MARGIN (OPM), POR ESTO MISMO PODEMOS INFERIR QUE UNA MEDIDA MAS REPRESENTATIVA DE LOS DATOS SERIA LA MEDIANA EN LUGAR DE LA MEDIA

In [48]:
```
px.histogram(df_clean, x = 'Book_to_Market_ratio')
```

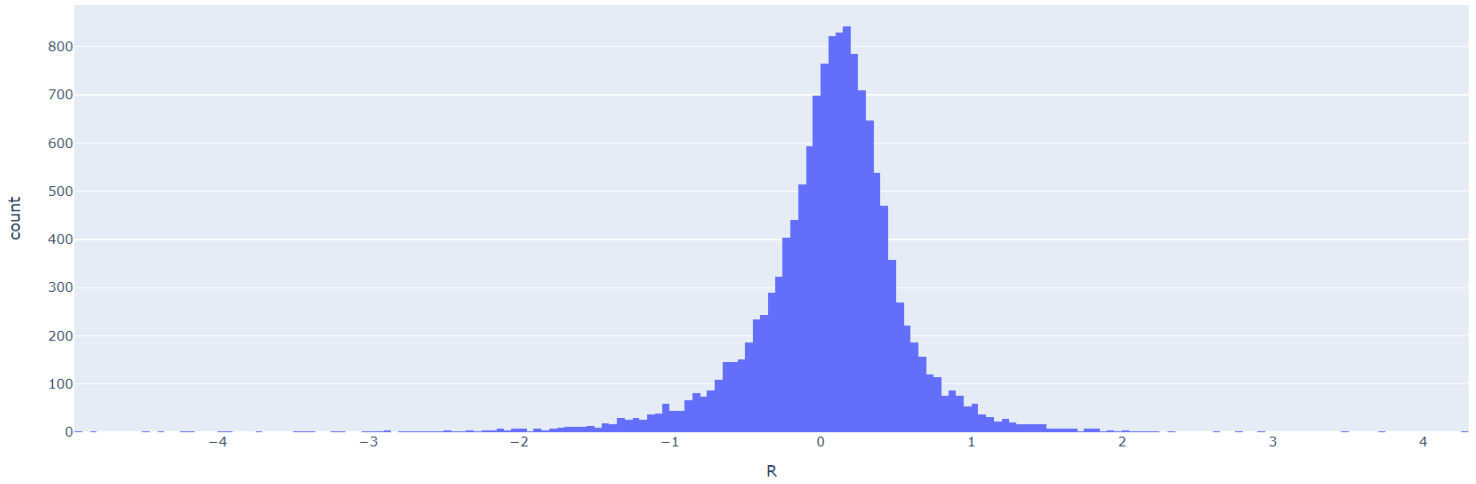


EN ESTE TERCER HISTOGRAMA VEMOS UNA DISTRIBUCION MAS EQUITATIVA DE LOS DATOS, SIN TENER CAMBIOS TAN BRUSCOS. AL IGUAL QUE EN EL RESTO DE VARIABLES, LOS DATOS SE ENCUENTRAN EN SU MAYORIA EN 0, AUQUE SE PUEDE APRECIAR UNA CAMPANA DE GAUSS CON DATOS MEJOR DEISTRIBUIDOS QUE EN LAS PRIMERAS DOS, POR ESTO MISMO PODRIAMOS USAR TANTO MEDIA COMO MEDIANA COMO MEDIDA DESCRIPTIVA, PERO AL TENER TANTOS DATOS EN 0 CONSIDERO MEJOR LA MEDIANA

In [49]:
```
df_clean['R'] = np.log(df_clean.groupby(['firm'])['adjprice'].shift(-1)) - np.log(df_clean.groupby(['firm'])['adjprice'].shift(3))
```

In [50]:
```
df_clean_mask = df_clean['R'] >= -1000000
df_clean = df_clean[df_clean_mask]
```

In [51]:
```
px.histogram(df_clean, x = 'R')
```



EN CUANTO A LA VARIABLE DE LA R, AL IGUAL QUE EN EL BOOK TO MARKER RATIO, CONTAMOS CON UNA DISTRIBUCION MAS NORMAL DE LOS DATOS, YA QUE SE PUEDE VEWR COMO TODOS SE ENCUENTRAN ENTRE -5 Y 5 Y EN SU MAYORIA ESTAN CENTRALIZADOS EN EL 0, GENERANDO UNA CAMPANA DE GAUSS

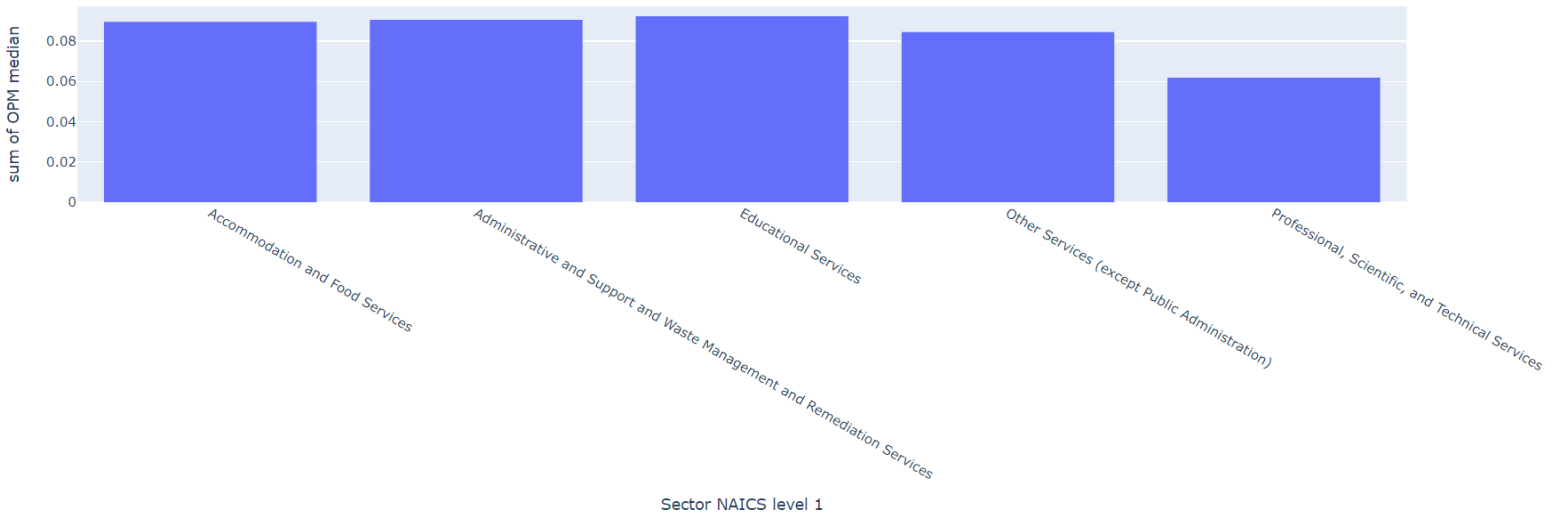
In [52]:
```
df_services_clean = df_clean.groupby('Sector NAICS\nlevel 1')['OPM'].median().to_frame()
df_services_clean['OPM mean'] = df_clean.groupby('Sector NAICS\nlevel 1')['OPM'].mean()
df_services_clean['EPSP median'] = df_clean.groupby('Sector NAICS\nlevel 1')['EPSP'].median()
df_services_clean['EPSP mean'] = df_clean.groupby('Sector NAICS\nlevel 1')['EPSP'].mean()
df_services_clean['Book to Market ratio median'] = df_clean.groupby('Sector NAICS\nlevel 1')['Book_to_Market_ratio'].median()
df_services_clean['Book to Market ratio mean'] = df_clean.groupby('Sector NAICS\nlevel 1')['Book_to_Market_ratio'].mean()
df_services_clean['R median'] = df_clean.groupby('Sector NAICS\nlevel 1')['R'].median()
df_services_clean['R mean'] = df_clean.groupby('Sector NAICS\nlevel 1')['R'].mean()
df_services_clean.reset_index(inplace=True)
df_services_clean.columns = df_services_clean.columns.str.replace('OPM', 'OPM median')
df_services_clean.columns = df_services_clean.columns.str.replace('OPM median mean', 'OPM mean')
df_services_clean.head()
```

Out[52]:

	Sector NAICS\nlevel 1	OPM median	OPM mean	EPSP median	EPSP mean	Book to Market ratio median	Book to Market ratio mean	R median	R mean
0	Accommodation and Food Services	0.089674	-8.916966	0.009384	-0.000764	0.257698	0.376347	0.111632	0.089202
1	Administrative and Support and Waste Managemen...	0.090587	-0.366443	0.008503	-0.002688	0.297543	0.364688	0.124652	0.082372
2	Educational Services	0.092409	0.082680	0.008792	-0.005623	0.517745	0.602595	0.020446	0.011288
3	Other Services (except Public Administration)	0.084592	-1.836236	0.009281	-0.020158	0.369235	0.369043	0.053908	-0.003335
4	Professional, Scientific, and Technical Services	0.061885	-6.159392	0.005856	-0.022095	0.312910	0.386048	0.077051	0.022697

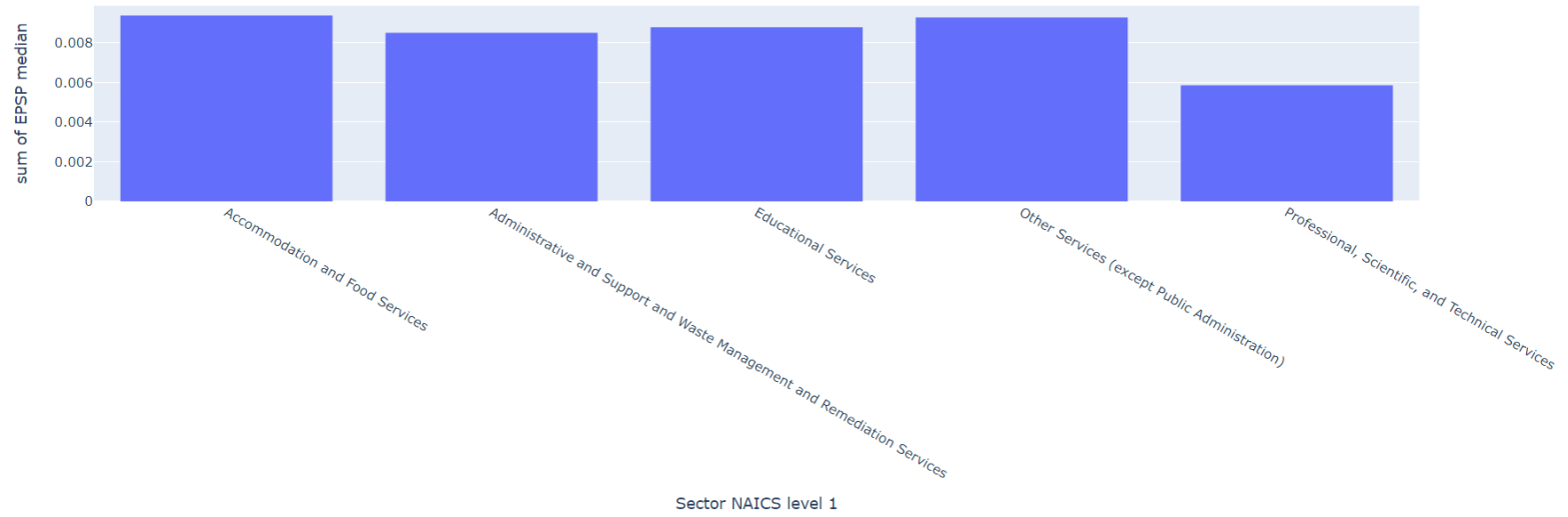
COMO PODEMOS VER , LA MEDIA Y LA MEDIANA DE LAS VARIABLES "OPM" Y "EPSP", SON MUY DIFERENTES, ESTO SE DEBE A QUE EXISTE UNA INCLINACION HACIA LA DERECHA EN LOS DATOS, DONDE TENEMOS A LA MAYORIA DE EMPRESAS CON VALORES MUY CERCANOS A 0, PERO EXISTEN ALGUNAS CON VALORES NEGATIVOS, LO CUAL GENERA ESTA DIFERENCIA, POR LO QUE TOMAREMOS LA MEDIANA COMO MEDIDA ESTANDAR EN LUGAR DE LA MEDIA

In [53]:
```
px.histogram(df_services_clean, x = 'Sector NAICS\nlevel 1', y = 'OPM median')
```



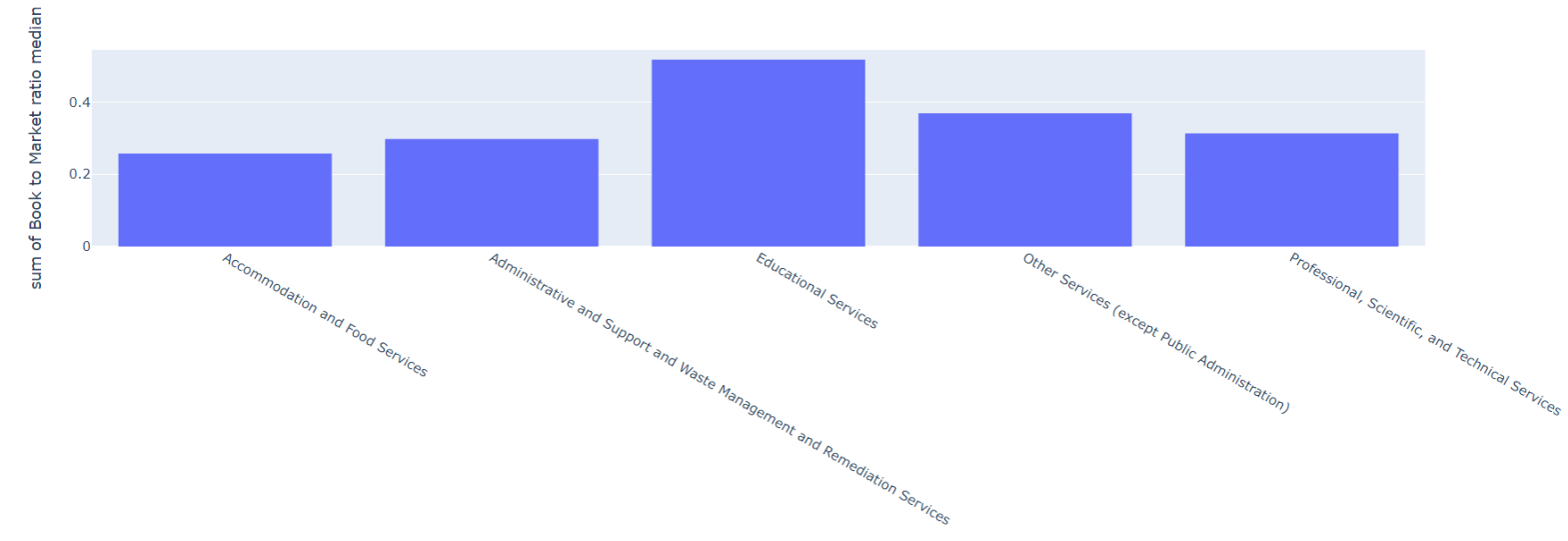
EN ESTOS HISTOGRAMAS, A DIFERENCIA DE LOS ANTERIORES, ESTAMOS GRAFICANDO SEGUN CAGA SECTOR, Y SEGUN LA MEDIANA DE LOS DATOS POR LO MENCIONADO ANTERIOR MENTE. PODEMOS VER QUE CASI TODOS LOS SECTORES, A ESEPCION DEL PROFESSIONAL, SCIENTIFIC AND TECHNICAL SERVICES, CUENTAN CON UN OPM MAYOR A 0.08, ESTO QUIERE DECIR QUE ESTA INDUSTRIA CUENTA CON UN PORCENTAJE DE GANANCIA MEDIA ENTES DE IMPUESTOS DE 0.08

In [54]:
```
px.histogram(df_services_clean, x = 'Sector NAICS\nlevel 1', y = 'EPSP median')
```



AL IGAL QUE EN EL HISTOGRAMA ANTERIOR, PODEMOS VER QUE EL SECTOR CON MENOR EPSP SERIA EL DE SCIENTIFIC AND TECHNICAL SERVICES, Y QUE LA MEDIANA DE CADA SECTOR ESTA CERCANA A 0.08

In [55]:
```
px.histogram(df_services_clean, x = 'Sector NAICS\nlevel 1', y = 'Book to Market ratio median')
```



FINALMENTE, EN ESTE GRAFICO SE ENCUENTRA MAYOR VARIACION ENTRE SECTORES, TENIENDO EDUCATIONAL SERVICES COMO EL DE MAYOR MEDIANA Y ACCOMODATION AND FOOD SERVICES COMO EL DE MENOR

Design and run a multiple regression model to examine whether your selected factors and earnings per share deflated by price can explain/predict annual stock returns. You have to control for industry and firm size. To control for these variables you have to include them as extra independent variables in the model

In [56]:
```
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

CREAMOS UN DATAFRAME UNICAMENTE CON LOS DATOS EMPLEADOS EN EL MODELO

In [57]:
```
df_modelo = df_clean[['Book_to_Market_ratio', 'OPM', 'EPSP', 'isSmall', 'isMedium', 'R']]
```

In [58]:
```
from scipy.stats.mstats import winsorize
df_modelo["EPSP"] = winsorize(df_modelo["EPSP"], limits=[0.0001, 0.02])
df_modelo["Book_to_Market_ratio"] = winsorize(df_modelo["Book_to_Market_ratio"], limits=[0.0001, 0.02])
df_modelo["OPM"] = winsorize(df_modelo["OPM"], limits=[0.0001, 0.02])
df_modelo["R"] = winsorize(df_modelo["R"], limits=[0.0001, 0.02])
```

WINZORIZAMOS LOS DATOS PARA MODIFICAR OUTLIERS TANTO DE ARRIBA COMO ABAJO, EN ESTE CASO ELEGI EL 2% DE LOS DATOS QUE SE ENCUENTTRAN POR ARRIBA Y 0.01% DE LOS QUE SE ENCUENTRAN POR DEBAJO, ESTO YA QUE TENEMOS MUCHOS OUTLIERS CON DATOS MUCHO MAYORES A LA MEDIANA Y POCO CON DATOS MUCHO MENORES

In [59]:
```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif_data = pd.DataFrame()
vif_data["feature"] = df_modelo.columns

# calculating VIF for each feature
vif_data["VIF"] = [variance_inflation_factor(df_modelo.values, i)
                          for i in range(len(df_modelo.columns))]

print(vif_data)
```

```
                feature       VIF
0  Book_to_Market_ratio  1.505610
1                   OPM  1.006534
2                  EPSP  1.118257
3               isSmall  1.383779
4              isMedium  1.152708
5                     R  1.103508
```

COMO PODEMOS VER, AL TENER UN VIF DE MENOS DE 1.5 EN TODAS LAS VARIABLES, NO EXISTE MULTICOLIENARIDAD EN LOS DATOS, POR LO QUE NO SE NECESITA GENERAR CAMBIOS EN ESTOS

In [60]:
```
mod = smf.ols('R ~ Book_to_Market_ratio + OPM + EPSP + isSmall + isMedium', data = df_modelo).fit()

print(mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      R   R-squared:                       0.109
Model:                            OLS   Adj. R-squared:                  0.109
Method:                 Least Squares   F-statistic:                     336.9
Date:                Tue, 13 Sep 2022   Prob (F-statistic):               0.00
Time:                        19:01:06   Log-Likelihood:                -8905.9
No. Observations:               13761   AIC:                         1.782e+04
Df Residuals:                   13755   BIC:                         1.787e+04
Df Model:                           5
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                  0.1383      0.007     19.441      0.000       0.124       0.152
Book_to_Market_ratio      -0.0769      0.008     -9.762      0.000      -0.092      -0.061
OPM                    6.534e-05   3.84e-05      1.701      0.089   -9.97e-06       0.000
EPSP                       1.0321      0.030     34.533      0.000       0.973       1.091
isSmall                   -0.1198      0.010    -11.817      0.000      -0.140      -0.100
isMedium                  -0.0484      0.010     -5.013      0.000      -0.067      -0.029
==============================================================================
Omnibus:                     3161.966   Durbin-Watson:                   0.771
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            21128.353
Skew:                          -0.937   Prob(JB):                         0.00
Kurtosis:                       8.774   Cond. No.                         781.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Interpret your model

AL CORRER EL MODELO, PODEMOS VER QUE LA VARIABLE BOOK TO MARKET RATIO, TIENE UNA RELACION NEGATIVA CON NUESTRA R, CON UNA PENDIENTE DE -0.0769, ESTO QUIERE DECIR QUE AL INCREMENTAR 1 EN EL BOOK TO MARKET RATIO, NUESTRO RETORNO DE STOCK DISMINUYE EN 0.0769, ESTA VARIABLE ES SIGNIFICATIVA, DEBIDO A QUE CUENTA CON UNA FIDELIDAD DE 99.999% DEBIDO AL P_VALUE DE 0.000.

EN CUANTO A LA OPM, ESTA TIENE UNA RELACION POSITIVA, AUNQUE MUY PEQUEÑA, YA QUE POR CADA AUMENTO EN OPM, EL RETORNO DE STOCK AUMENTARA UNICAMENTE EN 0.00006, DEBIDO A SU T_VALUE DE 1.701, PODEMOS DECIR QUE ESTA VARIABLE ES SIGNIFICATIVA EN UN 99.911%.

LA TERCERA VARIABLE ES EPSP, LA CUAL IGUALMENTE CUENTA CON UNA RELACION POSITIVA DE 1.0321, QUE AL IGUAL QUE EN LAS VARIABLES ANTERIORES, ESTO QUIERE DECIR QUE POR CADA AUMENTO EN EPSP, LA R AUMENTA EN 1.0321, Y ESTA VARIABLE VUELVE A SER SIGNIFICATIVA, DEBIDO A SU T_VALUE DE 34, LO CUAL NOS ASEGURA UN 99.99999% DE FIDELIDAD.

FINALMENTE, AL DIVIDIR LAS EMPRESAS EN CHICA MEDIANA Y GRANDE, PODEMOS VER QUE SI NUESTRA EMPRESA ES GRANDE, TENDREMOS MAYORES GANANCIAS A DIFERENCIA DE SI ES MEDIANA O LA CHICA, ESTO DEBIDO A QUE CUANDO TENEMOS UNA EMPRESA GRANDE, EMPEZAMOS CON UN 0.1383 DE RETORNOS DE STOCK, EN UNA MEDIANA ESTO DISMINUIRIA A 0.0899 Y EN UNA PEQUEÑA, EMPEZARIAMOS CON 0.0185. TODAS ESTAS VARIABLES SON SIGNIFICATIVAS, YA QUE CUENTAN CON UNA FIDELIDAD DE 99.99%.

ESTE MODELO, REALMENTE SOLO ES ASERTADO EN UN 10.9% DE LOS CASOS, ESTO ES DEBIDO A NUESTRA R^2 DE 0.109, PERO ES UN RESULTADO ESPERABLE, DEBIDO A LA VARIABILIDAD DE LOS RETORNOS DE STOCK DE LAS EMPRESAS, LO CUAL HACE REALMENTE COMPLICADO PODER LLEGAR A PREDECIR A FUTURO SUS VALORES.

BIG COMPANIES

R = 0.1383 - 0.0769(Book_to_market_ratio) + 0.00006(OPM) + 1.0321(EPSP)

MEDIUM COMPANIES

R = 0.0899 - 0.0769(Book_to_market_ratio) + 0.00006(OPM) + 1.0321(EPSP)

SMALL COMPANIES

R = 0.0185 -0.0769(Book_to_market_ratio) + 0.00006(OPM) + 1.0321(EPSP)