DATA 602 PROJECT PROPOSAL

Covid Rates and Income

Research Question

Is there a relationship between Covid rates (cases, hospitalizations, death) and Income?

Justification

This question has relevance across industries. Understanding Covid susceptibility allows for better risk assessment and contingency planning for future public health events.

Data Sources

This analysis will focus on three datasets from two data sources:

- IRS Data by Zip Code - 2019 (source: US Dept of Treasury)
- Provisional COVID-19 Death Counts in the United States by County (source: CDC)
- United States COVID-19 Community Levels by County (source: CDC)

Libraries

This project will utilize the following libraries:

- Pandas
- NumPy
- Matplotlib
- Seaborn
- Additional libraries as needed

EDA & Summary Statistics

IRS DATA

In [25]:
```
irs.head()
```

Out[25]:

	STATEFIPS	STATE	zipcode	agi_stub	N1	mars1	MARS2	MARS4	ELF	CPREP	...	N85300	A85300	N11901	A11901	N119
0	1	AL	0	1	778210.0	491030.0	84770.0	189600.0	712890.0	30670.0	...	0.0	0.0	62720.0	51936.0	67186
1	1	AL	0	2	525940.0	247140.0	123910.0	139860.0	481760.0	18960.0	...	0.0	0.0	85860.0	122569.0	43802
2	1	AL	0	3	285700.0	105140.0	128140.0	44560.0	260570.0	10670.0	...	0.0	0.0	73980.0	154932.0	21204
3	1	AL	0	4	179070.0	38820.0	123110.0	13740.0	164300.0	5020.0	...	0.0	0.0	51330.0	139065.0	12685
4	1	AL	0	5	257010.0	28180.0	216740.0	7150.0	236850.0	8400.0	...	90.0	141.0	104290.0	460071.0	15279

5 rows × 152 columns

In [26]:
```
irs.describe()
```

Out[26]:

	STATEFIPS	zipcode	agi_stub	N1	mars1	MARS2	MARS4	ELF	CPREP	
count	166159.000000	166159.000000	166159.000000	1.661590e+05	1.661590e+05	1.661590e+05	166159.00000	1.661590e+05	166159.000000	1.6615
mean	29.666885	48859.485553	3.499949	1.860508e+03	9.127834e+02	6.478571e+02	258.03676	1.689549e+03	75.771821	9.521
std	15.121486	27167.679271	1.707871	3.722335e+04	2.224999e+04	1.200080e+04	6336.06430	3.347049e+04	1866.726495	1.936
min	1.000000	0.000000	1.000000	0.000000e+00	0.000000e+00	0.000000e+00	0.00000	0.000000e+00	0.000000	0.000
25%	18.000000	27020.000000	2.000000	7.000000e+01	0.000000e+00	4.000000e+01	0.00000	7.000000e+01	0.000000	5.000
50%	29.000000	48843.000000	3.000000	2.600000e+02	8.000000e+01	1.100000e+02	30.00000	2.400000e+02	0.000000	1.500
75%	42.000000	70652.500000	5.000000	1.080000e+03	3.900000e+02	3.800000e+02	100.00000	9.900000e+02	40.000000	5.600
max	56.000000	99999.000000	6.000000	5.506120e+06	4.069770e+06	1.818210e+06	945490.00000	4.827070e+06	338290.000000	3.022

8 rows × 151 columns

In [27]:
```
irs.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 166159 entries, 0 to 166158
Columns: 152 entries, STATEFIPS to A12000
dtypes: float64(148), int64(3), object(1)
memory usage: 192.7+ MB
```

The key for this data has been loaded into the project GitHub repository. The original IRS dataset contains over 166k observations across 152 columns, and must be read in locally as it is 192.7mb in size. The data will transformed to its necessary components for this analysis, written to .csv, and uploaded to the project GitHub repository.

CDC Community Levels By County

In [24]:

```
cdc_comm.head()
```

Out[24]:

	county	county_fips	state	county_population	health_service_area_number	health_service_area	health_service_area_population	covid_inpatie
0	Lincoln County	55069	Wisconsin	27593.0	282	Marathon (Wausau), WI - Wood, WI	291401.0	
1	Manitowoc County	55071	Wisconsin	78981.0	355	Sheboygan (Sheboygan), WI - Manitowoc, WI	244410.0	
2	Marathon County	55073	Wisconsin	135692.0	282	Marathon (Wausau), WI - Wood, WI	291401.0	
3	Monroe County	55081	Wisconsin	46253.0	290	La Crosse (La Crosse), WI - Monroe, WI	257027.0	
4	Portage County	55097	Wisconsin	70772.0	400	Portage, WI	70772.0	

In [28]:

```
cdc_comm.describe()
```

Out[28]:

	county_fips	county_population	health_service_area_number	health_service_area_population	covid_inpatient_bed_utilization	covid_hospital_adr
count	112836.00000	1.128350e+05	112836.000000	1.128290e+05	112648.00000	
mean	31438.02789	1.029200e+05	400.462033	5.808604e+05	3.25508	
std	16331.50567	3.293638e+05	243.444960	9.952625e+05	2.66225	
min	1001.00000	8.600000e+01	1.000000	2.274000e+03	0.00000	
25%	19033.00000	1.113100e+04	186.000000	9.021200e+04	1.30000	
50%	30027.00000	2.611800e+04	409.000000	2.249140e+05	2.80000	
75%	46111.00000	6.721500e+04	587.000000	5.545570e+05	4.50000	
max	78000.00000	1.003911e+07	905.000000	1.321480e+07	36.00000	

In [29]:

```
cdc_comm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112836 entries, 0 to 112835
Data columns (total 12 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   county                            112836 non-null  object
 1   county_fips                       112836 non-null  int64
 2   state                             112836 non-null  object
 3   county_population                 112835 non-null  float64
 4   health_service_area_number        112836 non-null  int64
 5   health_service_area               112836 non-null  object
 6   health_service_area_population    112829 non-null  float64
 7   covid_inpatient_bed_utilization   112648 non-null  float64
 8   covid_hospital_admissions_per_100k 112778 non-null  float64
 9   covid_cases_per_100k              112836 non-null  float64
 10  covid-19_community_level          112782 non-null  object
 11  date_updated                      112836 non-null  object
dtypes: float64(5), int64(2), object(5)
memory usage: 10.3+ MB
```

This dataset contains over 112k observations across 12 columns containing information about Covid-19 cases and hospitalizations.

CDC Provisional Death Counts By County

In [22]:

```
cdc_prov.head()
```

Out[22]:

	Date as of	Start Date	End Date	State	County name	FIPS County Code	Urban Rural Code	Deaths involving COVID-19	Deaths from All Causes	Footnote
0	10/19/2022	01/01/2020	10/15/2022	AK	Aleutians East Borough	2013	Noncore	NaN	22.0	One or more data cells have counts between 1-9...
1	10/19/2022	01/01/2020	10/15/2022	AK	Anchorage Municipality	2020	Medium metro	734.0	7081.0	NaN
2	10/19/2022	01/01/2020	10/15/2022	AK	Bethel Census Area	2050	Noncore	39.0	317.0	NaN
3	10/19/2022	01/01/2020	10/15/2022	AK	Denali Borough	2068	Noncore	NaN	24.0	One or more data cells have counts between 1-9...
4	10/19/2022	01/01/2020	10/15/2022	AK	Dillingham Census Area	2070	Noncore	NaN	96.0	One or more data cells have counts between 1-9...

In [30]:

```
cdc_prov.describe()
```

Out[30]:

	FIPS County Code	Deaths involving COVID-19	Deaths from All Causes
count	3085.000000	2706.000000	3084.000000
mean	30357.156240	391.218404	3027.573281
std	15162.540083	1179.109564	8527.317813
min	1001.000000	10.000000	14.000000
25%	18175.000000	29.000000	304.000000
50%	29147.000000	75.000000	717.500000
75%	45075.000000	297.500000	2120.500000
max	56045.000000	31094.000000	223502.000000

In [31]:

```
cdc_prov.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3085 entries, 0 to 3084
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Date as of                3085 non-null   object 
 1   Start Date                3085 non-null   object 
 2   End Date                  3085 non-null   object 
 3   State                     3085 non-null   object 
 4   County name               3085 non-null   object 
 5   FIPS County Code          3085 non-null   int64  
 6   Urban Rural Code          3085 non-null   object 
 7   Deaths involving COVID-19  2706 non-null   float64
 8   Deaths from All Causes    3084 non-null   float64
 9   Footnote                  379 non-null    object 
dtypes: float64(2), int64(1), object(7)
memory usage: 241.1+ KB
```

This dataset contains 3085 observations across 10 columns, containing total Covid deaths by State and County.

Combining the Datasets

These datasets will be transformed and combined to create a master dataframe containing Zip code, Income, and Covid rate information.