```
# Run this cell to authenticate yourself to BigQuery.
from google.colab import auth
auth.authenticate_user()
project_id = "aerobic-star-218315"
```

```
# Some imports you will need
import pandas as pd
import altair as alt
```

```
# Initialize BiqQuery client
from google.cloud import bigquery
client = bigquery.Client(project=project_id)  # pass in your projectid
```

Analysis of dataset

We use the World Bank dataset for the project, specifically the world development indicators dataset. five tables. The first one is country_series_definitions, in which each row lists the country, series type, information about the country's demographics captured in the series data. The second table, country countries and country groups considered along with additional summarizing information such as inco geographical information, and census information. The third table, indicators_data, gives the actual va indicator for each country in a specified year. More information about the series in summarized in ser years covered by the data series is given in series_times.

indicators_data is the important table with the values of the keys. The data is stored as object, key, ye is (indicator_code, year). The objects are the countries, and additional information about the objects i countr_summary table. Information about the indicator code is found in series_summary table.

There are some pros and cons about the current design of the dataset. It is very easy to add new attri simply requires adding a new element description in the series_summary table and including the valu table. Adding new data every year is also made easy by this design. However, this design does make For example, deleting a certain indicator would require scanning the entire table, which is highly costl or sum over a certain indicator is also made harder by this design, since it requires scanning the whol extracting certain number of features for a certain country over many years requires the use of the sa which gets tedious for the user.

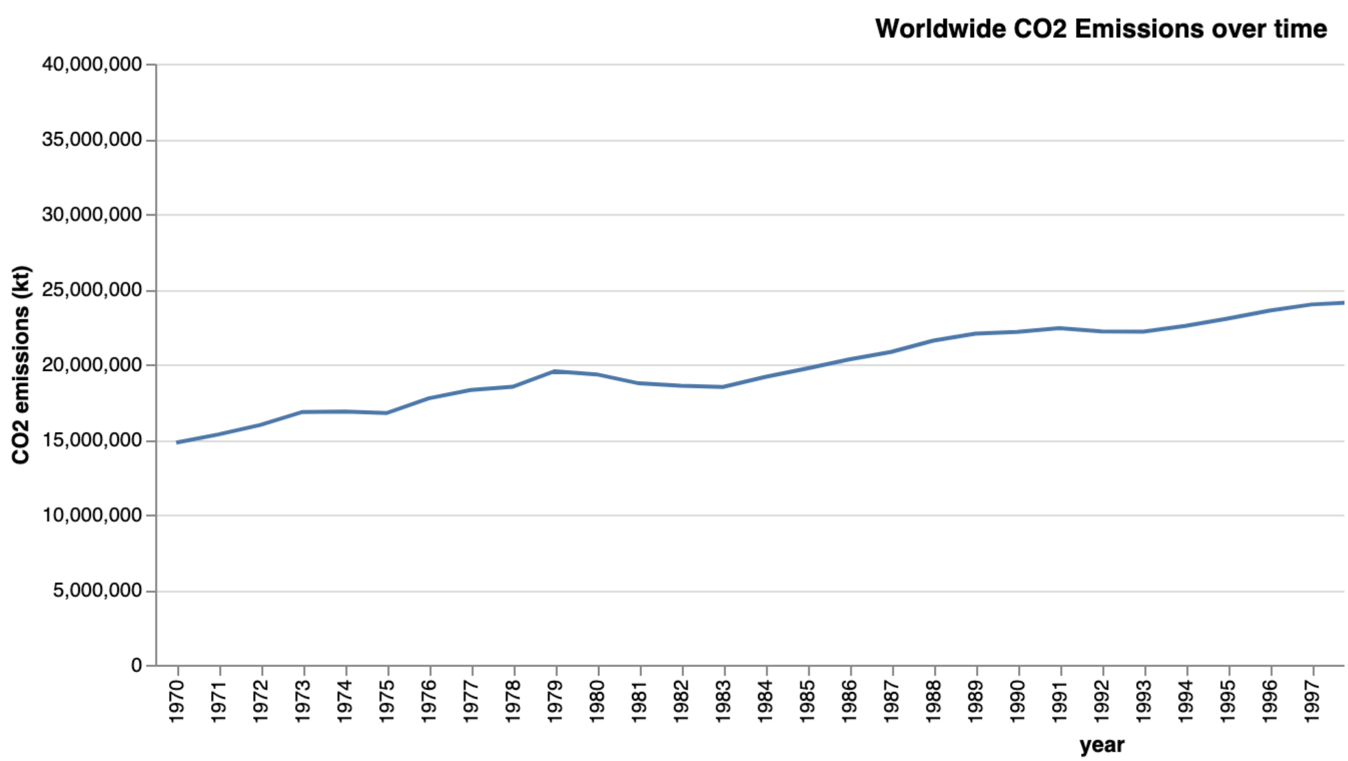
Data visualization

We start by performming some basic data visualization of CO2 trends think may be relevant to yearly emissions.

We are interested in predicting the per capita CO2 emissions of a country in a given year. First, let us emissions data.

```
%%bigquery --project $project_id p1

SELECT year, value, indicator_code, country_code
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data`
WHERE
  (country_code= "WLD" AND
  indicator_code= "EN.ATM.CO2E.KT")
```

```
alt.Chart(p1, title='Worldwide CO2 Emissions over time').mark_line().encode(
    x="year:N",
    y=alt.Y("value", axis=alt.Axis(title='CO2 emissions (kt)')))
```

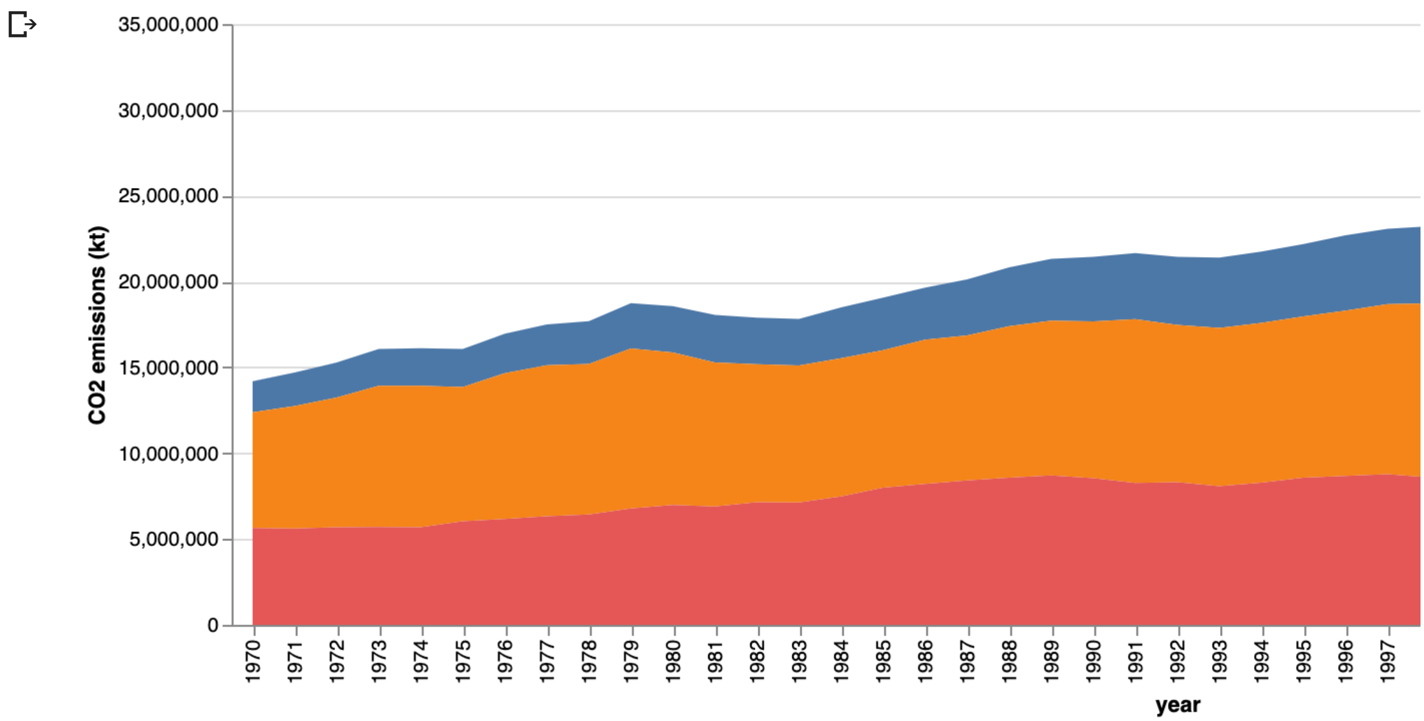


Export as SVG Export as PNG View Source Open in Vega Editor

Let us first get a big picture sense of the data and scale we are working with. Unsurprisingly, we see t emissions have increased over time. The graph above suggests that the growth is in fact acceleratin 5 decades, C02 emmisions have more than doubled, from 15M kt to over 35M kt.

```
%%bigquery --project $project_id p2

SELECT year, value, indicator_code, country_code
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data`
WHERE
  (country_code= "WLD" AND(
  indicator_code= "EN.ATM.CO2E.GF.KT" OR indicator_code= "EN.ATM.CO2E.SF.KT" OR indic
```

```
alt.Chart(p2).mark_area().encode(
    x="year:N",
    y=alt.Y("value", axis=alt.Axis(title='CO2 emissions (kt)'),stack="zero"),
    color="indicator_code")
```



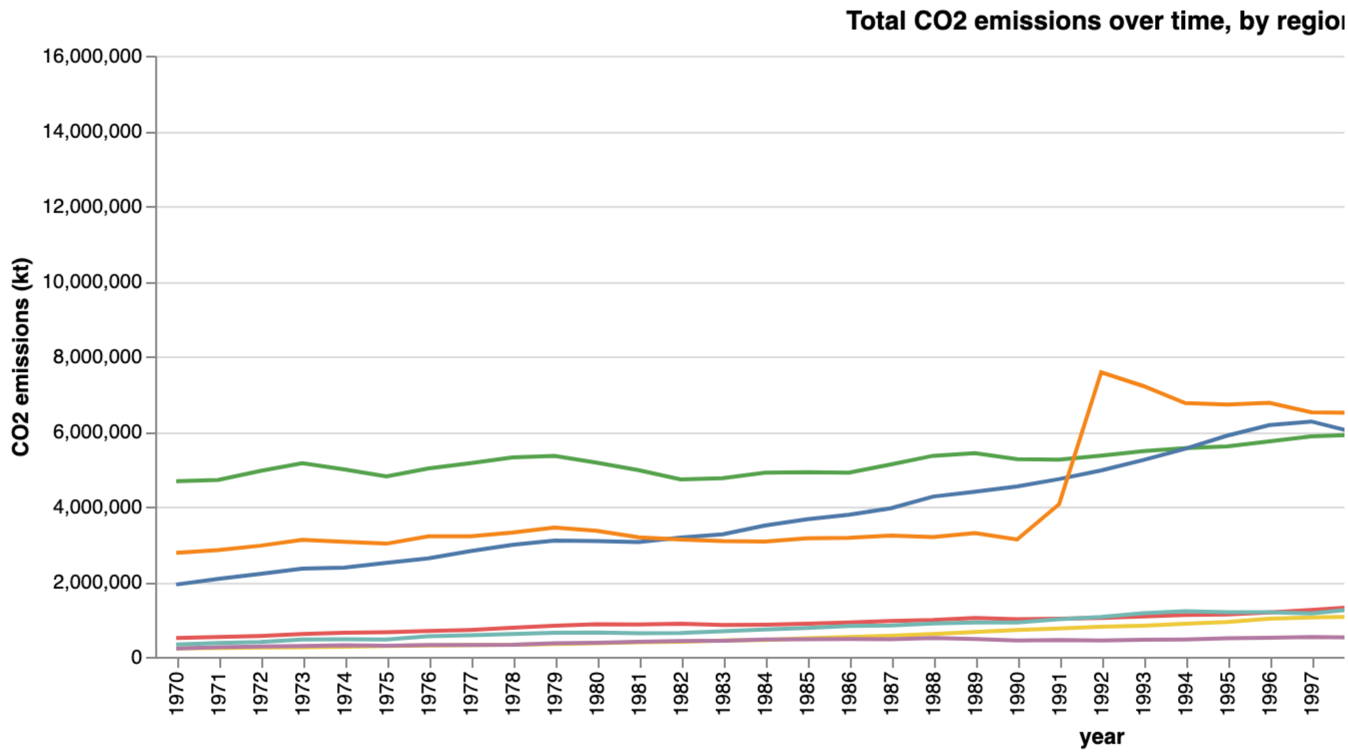
Export as SVG Export as PNG View Source Open in Vega Editor

We can see that over the past 5 decades, the fraction of CO2 emissions from solid fuels has increase from liquid fuels has decreased. The fraction from gaseous fuels remains the smallest.

```
%%bigquery --project $project_id p3


SELECT year, sum(value) as total_emissions, summary.region as region
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` indicators,
  `bigquery-public-data.world_bank_health_population.country_summary` summary

WHERE
```

```
  indicator_code= "EN.ATM.CO2E.KT" AND
  summary.country_code= indicators.country_code AND summary.region != ""
GROUP by year, region
```

```
alt.Chart(p3, title='Total CO2 emissions over time, by region').mark_line().encode(
    x="year:N",
    y=alt.Y("total_emissions", axis=alt.Axis(title='CO2 emissions (kt)')),
    color= "region")
```

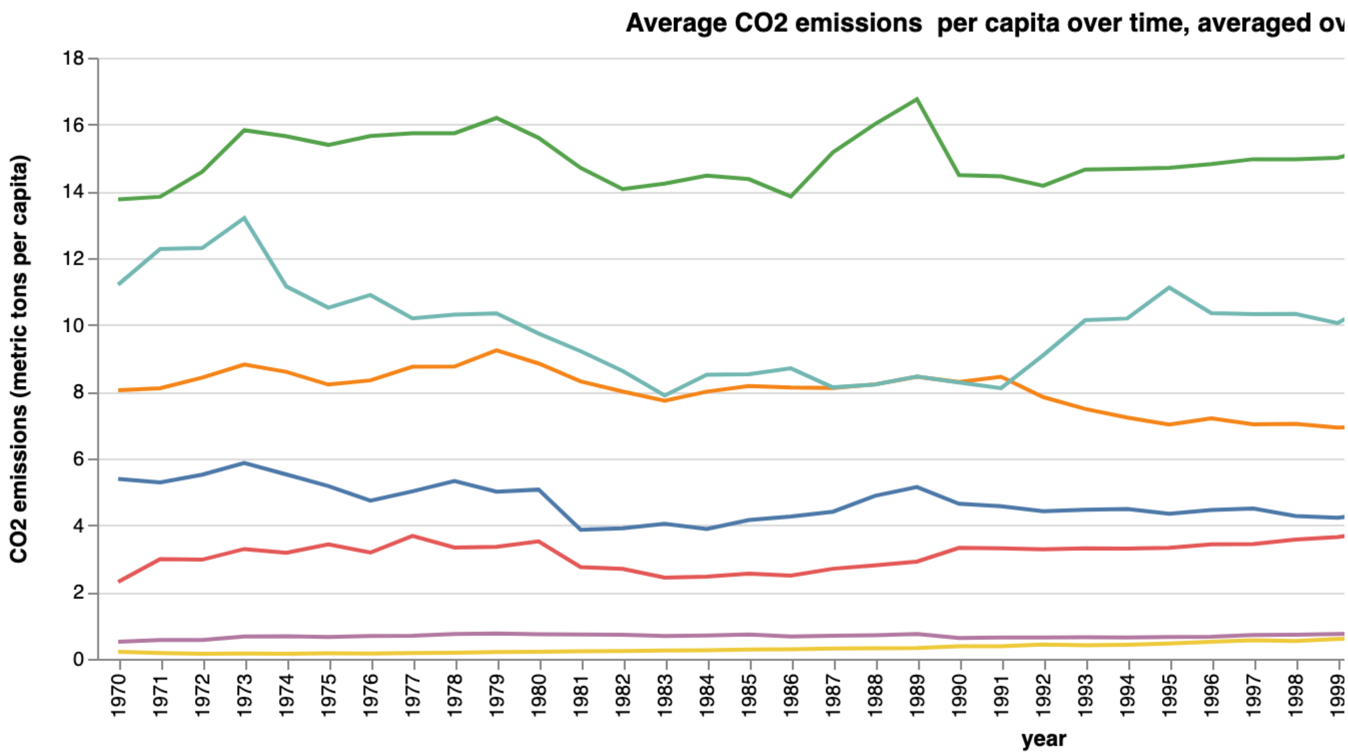


Export as SVG Export as PNG View Source Open in Vega Editor

Raw CO2 emissions are dominated by East Asia & Pacific, Europe & Central Asia, and North America. & Pacific has the highest rate of growth, with emissions increasing nearly 7 fold over roughly the past constrast, emissions in North America have only increased by about 5M kt. We note a large spike in E emmisions in 1992. We explore this peculiarity later below.

```
%%bigquery --project $project_id p4


SELECT year, avg(value) as total_emissions, summary.region as region
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` indicators,
  `bigquery-public-data.world_bank_health_population.country_summary` summary

WHERE
  indicator_code= "EN.ATM.CO2E.PC" AND
  summary.country_code= indicators.country_code AND summary.region != ""
GROUP by year, region
```

```
alt.Chart(p4, title='Average CO2 emissions  per capita over time, averaged over regio
    x="year:N",
    y=alt.Y("total_emissions", axis=alt.Axis(title='CO2 emissions (metric tons per ca
    color= "region")
```



Export as SVG Export as PNG View Source Open in Vega Editor

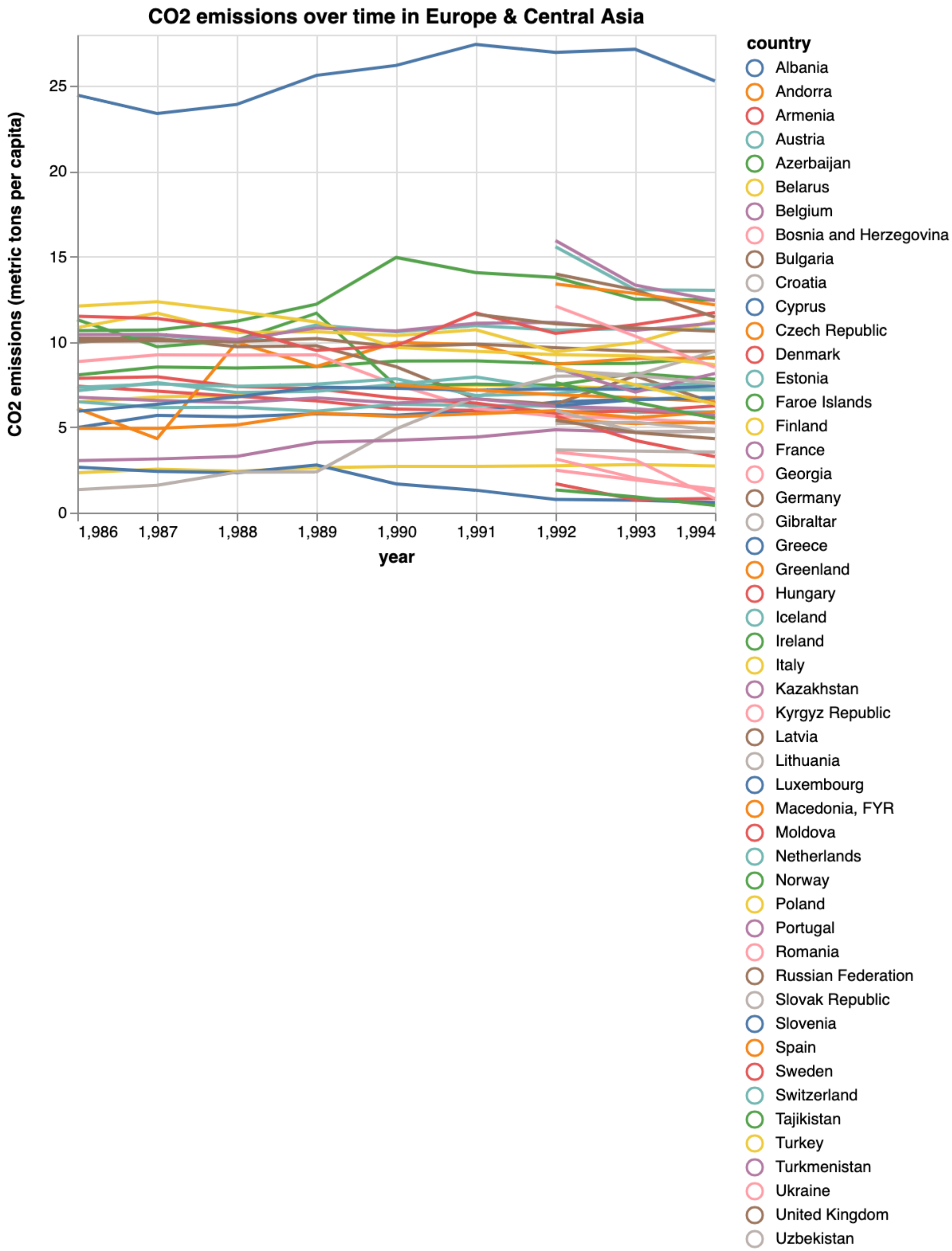
Looking at the average CO2 emissions per capita in each region per year reveals an interesting story. America (dominated by the United States and Canada) has the highest average emissions, followed b North Africa and then Europe & Central Asia. Wheras previous graphs showed that the East Asia & Pa production of raw emissions, the region falls in ranking when the emissions are normalized by popula per capita emissions. The graph suggests that developed countries rich in oil/coal/natural gas reserv Canada) have higher emissions.

Let's take a quick detour to dig deeper into the spike in the Europe & Central Asia data.

```
%%bigquery --project $project_id p5


SELECT year, value , indicators.country_name as country
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` indicators,
  `bigquery-public-data.world_bank_health_population.country_summary` summary

WHERE
```

```
  indicator_code="EN.ATM.CO2E.PC" AND
  summary.country_code= indicators.country_code AND summary.region = "Europe & Centra
  AND year>1985 AND year<1995
```

```
alt.Chart(p5, title='CO2 emissions over time in Europe & Central Asia').mark_line().e
    tooltip=['country'],
    x="year",
    y=alt.Y("value", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
    color= "country")
```



Export as SVG   Export as PNG   View Source   Open in Vega Editor

Above, we zoom into the decade 1985-1995. The year 1992 stands out. Why is that? The graph above
countries geographically located in Europe and Central Asia became members of the World Bank Gro
ones (ie, countries with high CO2 emissions) include Kazakhstan, Estonia, and Russia. The timing ma

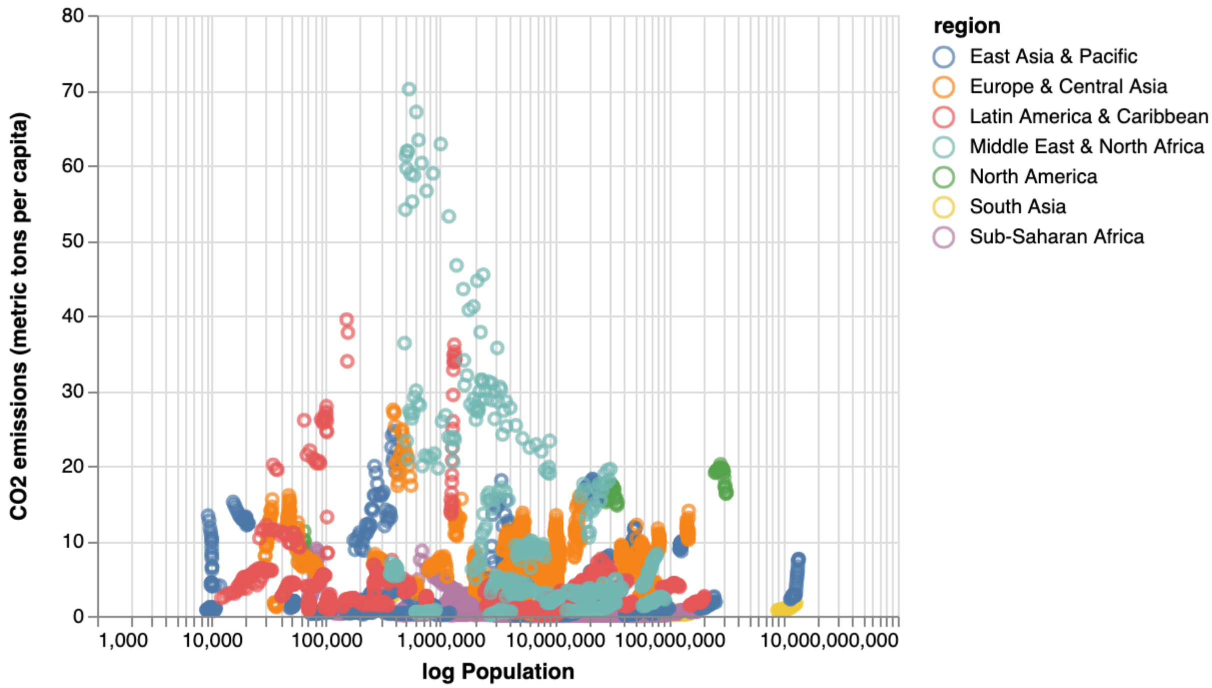
December 25, 1991, the USSR was dissolved into 15 post-Soviet states, and many became member c

Next, we explore some features that we think may play a role in determ emissions.

```
%%bigquery --project $project_id p6
```

```
SELECT  pop.value as population, pop.country_name, co2.value as co2_pc, summary.regio
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` pop,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
   pop.indicator_code= "SP.POP.TOTL" AND
   pop.country_code= co2.country_code AND
   pop.year= co2.year AND
   co2.indicator_code = "EN.ATM.CO2E.PC" AND
   summary.country_code= pop.country_code AND
   summary.country_code= co2.country_code AND
   summary.region != "" AND
   co2.year >1990 AND pop.year > 1990
```

```
# YOUR PLOT CODE HERE

slider = alt.binding_range(min=1990, max=2014, step=1)
select_year = alt.selection_single(name="year", fields=['year'], bind=slider)

alt.Chart(p6).mark_point().encode(
    x=alt.X("population", axis=alt.Axis(title='log Population'), scale=alt.Scale(type
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
    tooltip=['country_name'],
    color= alt.Color('region')
).add_selection(
    select_year).transform_filter(
    select_year
)
```

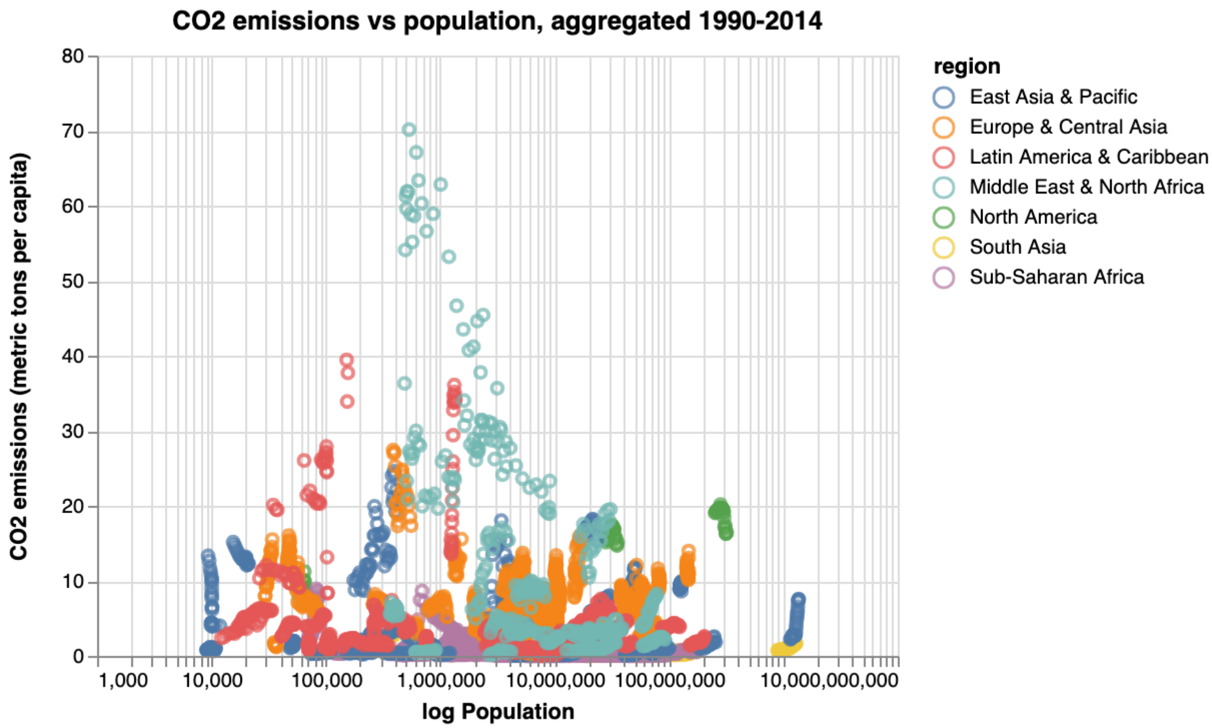


year_year

Export as SVG   Export as PNG   View Source   Open in Vega Editor

Above, we plot the CO2 emissions vs log population in a graph colored by region. A slider allows us to change year by year.

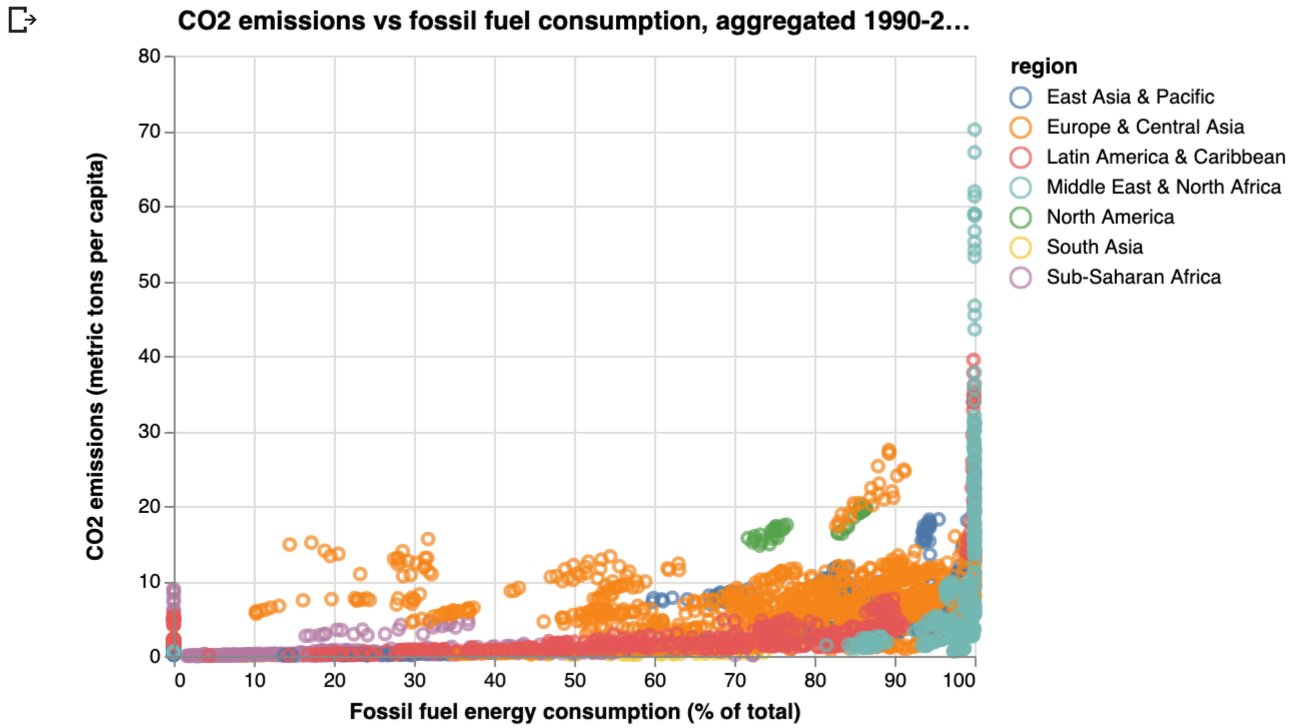
```
alt.Chart(p6, title='CO2 emissions vs population, aggregated 1990-2014').mark_point()
    tooltip=['country_name',"year"],
    x=alt.X("population", axis=alt.Axis(title='log Population'), scale=alt.Scale(type
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```



Export as SVG   Export as PNG   View Source   Open in Vega Editor

Above, we have a similar plot, but we aggregate all the data from 1990 to 2014. We see there is not a between the population of a country and amount of CO2 emissions. We see that the two most populc world, India and China, have fewer emissions that some of the Middle eastern countries with nearly 1

```
%%bigquery --project $project_id p7

SELECT  ff.value as fossil_fuel, ff.country_name, co2.value as co2_pc, summary.region
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` ff,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
   ff.indicator_code= "EG.USE.COMM.FO.ZS" AND
   ff.country_code= co2.country_code AND
   ff.year= co2.year AND
   co2.indicator_code = "EN.ATM.CO2E.PC" AND
   summary.country_code= ff.country_code AND
   summary.country_code= co2.country_code AND
   summary.region != "" AND
   co2.year >=1990 AND ff.year > 1990


alt.Chart(p7, title='CO2 emissions vs fossil fuel consumption, aggregated 1990-2014')
    tooltip=['country_name','year'],
    x=alt.X("fossil_fuel", axis=alt.Axis(title='Fossil fuel energy consumption (% of
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```

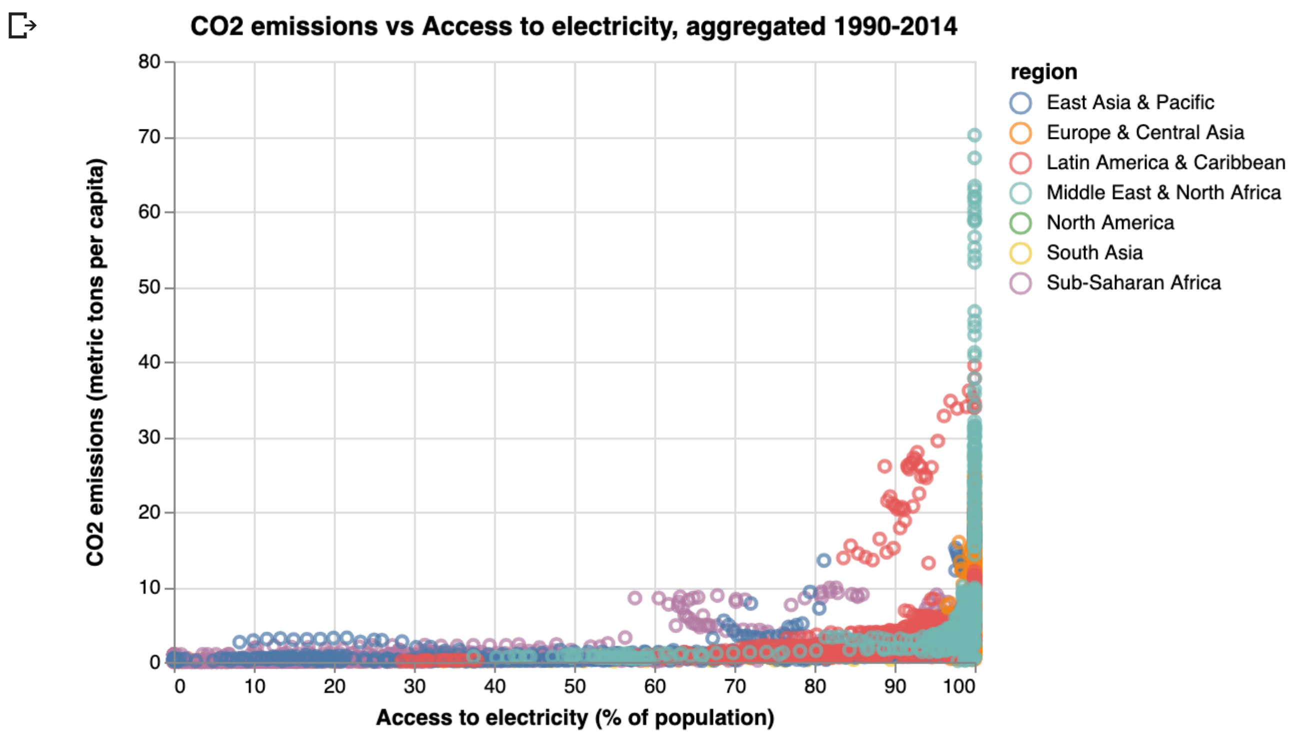


Export as SVG Export as PNG View Source Open in Vega Editor

Above, we look at the relationship between fossil fuel energy consumption and CO2 emissions. The g
that countries with higer fossil fuel energy consumption percentages tend to emit more CO2. Howeve
interesting exceptions. We see that many countries have 0% fossil fuel energy consumption, but have
emissions. Upon looking closer, we see that these countries include St. Kitts and Nevis, Equatorial Gu
Bahamas, etc. Interestingly, these are all islands. Additionally, many European countries, such as Icela
low fossil fuel energy consumption percentage, but high emissions.

```
%%bigquery --project $project_id p8


SELECT  access.value as access_elec, access.country_name, co2.value as co2_pc, summar
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` access,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
   access.indicator_code= "EG.ELC.ACCS.ZS" AND
   access.country_code= co2.country_code AND
   access.year= co2.year AND
   co2.indicator_code = "EN.ATM.CO2E.PC" AND
   summary.country_code= access.country_code AND
   summary.country_code= co2.country_code AND
   summary.region != "" AND
   co2.year >1990 AND access.year > 1990
```

```
alt.Chart(p8, title='CO2 emissions vs Access to electricity, aggregated 1990-2014').m
    tooltip=['country_name','year'],
    x=alt.X("access_elec", axis=alt.Axis(title='Access to electricity (% of populatio
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```



Export as SVG   Export as PNG   View Source   Open in Vega Editor

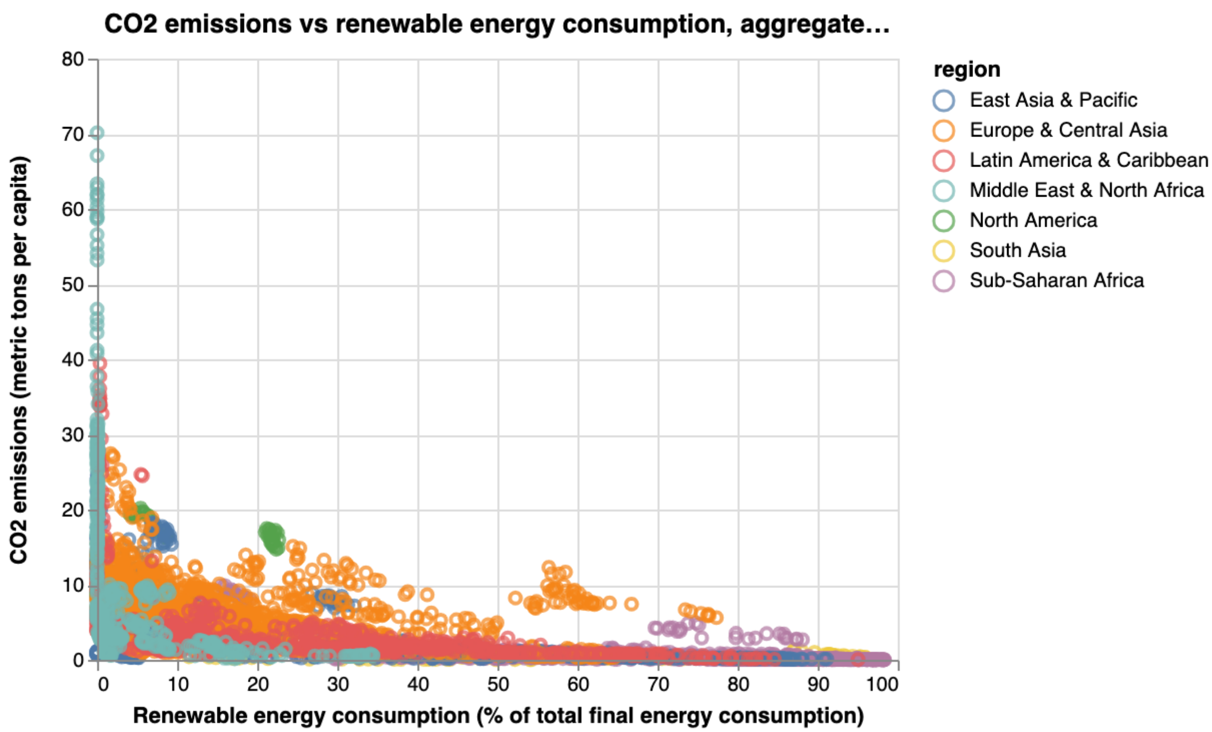
Above, we look at the relationship between CO2 emissions and the percentage of a country's populat electricity. The graph suggests that countries where less than 50% of the population has access to el per capita. In the highest emitting countries, a super majority of the population has access to electric

```
%%bigquery --project $project_id p9



SELECT  renew.value as renew_elec, renew.country_name, co2.value as co2_pc, summary.r
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` renew,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
   renew.indicator_code= "EG.FEC.RNEW.ZS" AND
   renew.country_code= co2.country_code AND
   renew.year= co2.year AND
   co2.indicator_code = "EN.ATM.CO2E.PC" AND
   summary.country_code= renew.country_code AND
   summary.country_code= co2.country_code AND
   summary.region != "" AND
```

```
  co2.year >1990 AND renew.year > 1990
```

```
;
```

```
alt.Chart(p9, title='CO2 emissions vs renewable energy consumption, aggregated 1990-2
    tooltip=['country_name',"year"],
    x=alt.X("renew_elec", axis=alt.Axis(title='Renewable energy consumption (% of tot
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```



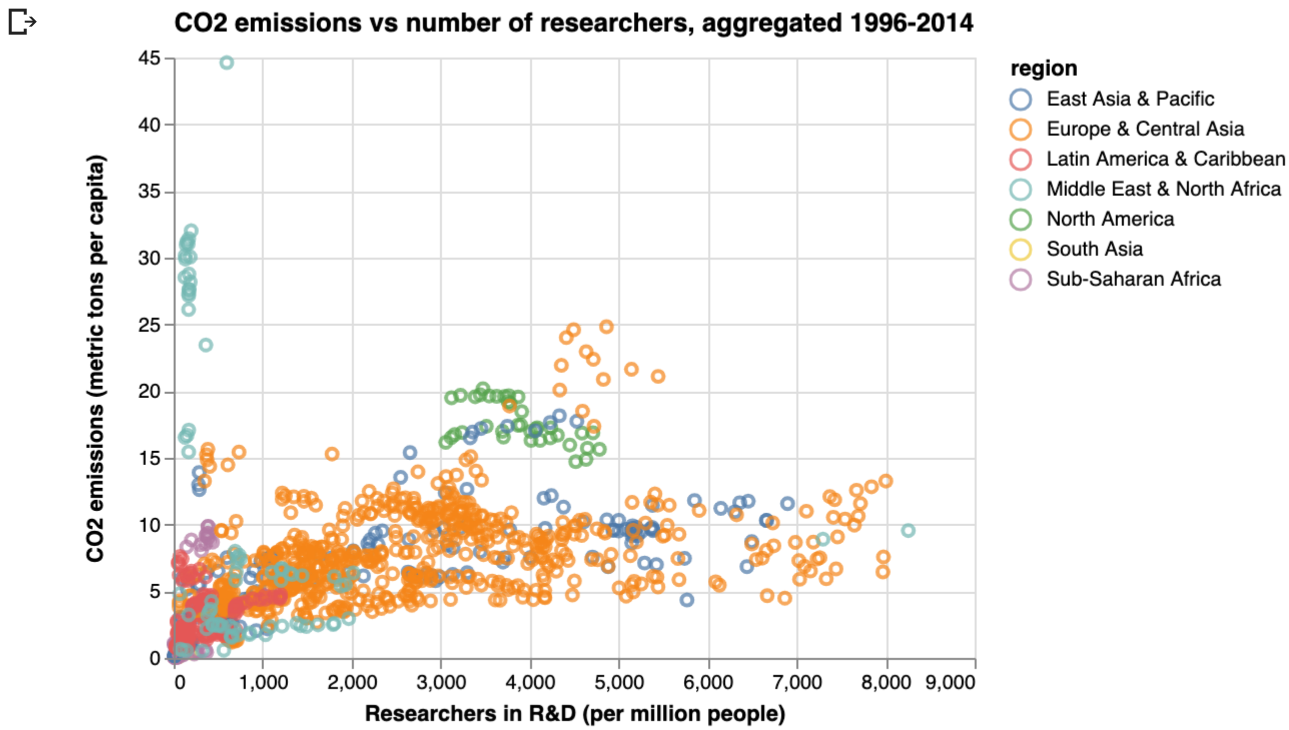
Export as SVG Export as PNG View Source Open in Vega Editor

We next study the role of renewable energy. Countries that have high renewable energy consumption lower emissions. Middle Eastern countries tend to have the lowest consumption percentage of renew countries with the highest emissions tend to be from the Middle East. Europe and North America alsc renewable energy consumption and has higher emissions than Latin American countries with compa consumption rates.

```
%%bigquery --project $project_id p10

SELECT  research.value as researchers, research.country_name, co2.value as co2_pc, su
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` research,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
```

```
    research.indicator_code= "SP.POP.SCIE.RD.P6" AND
    research.country_code= co2.country_code AND
    research.year= co2.year AND
    co2.indicator_code = "EN.ATM.CO2E.PC" AND
    summary.country_code= research.country_code AND
    summary.country_code= co2.country_code AND
    summary.region != ""
```

```
alt.Chart(p10, title='CO2 emissions vs number of researchers, aggregated 1996-2014').
    tooltip=['country_name',"year"],
    x=alt.X("researchers", axis=alt.Axis(title='Researchers in R&D (per million peopl
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```



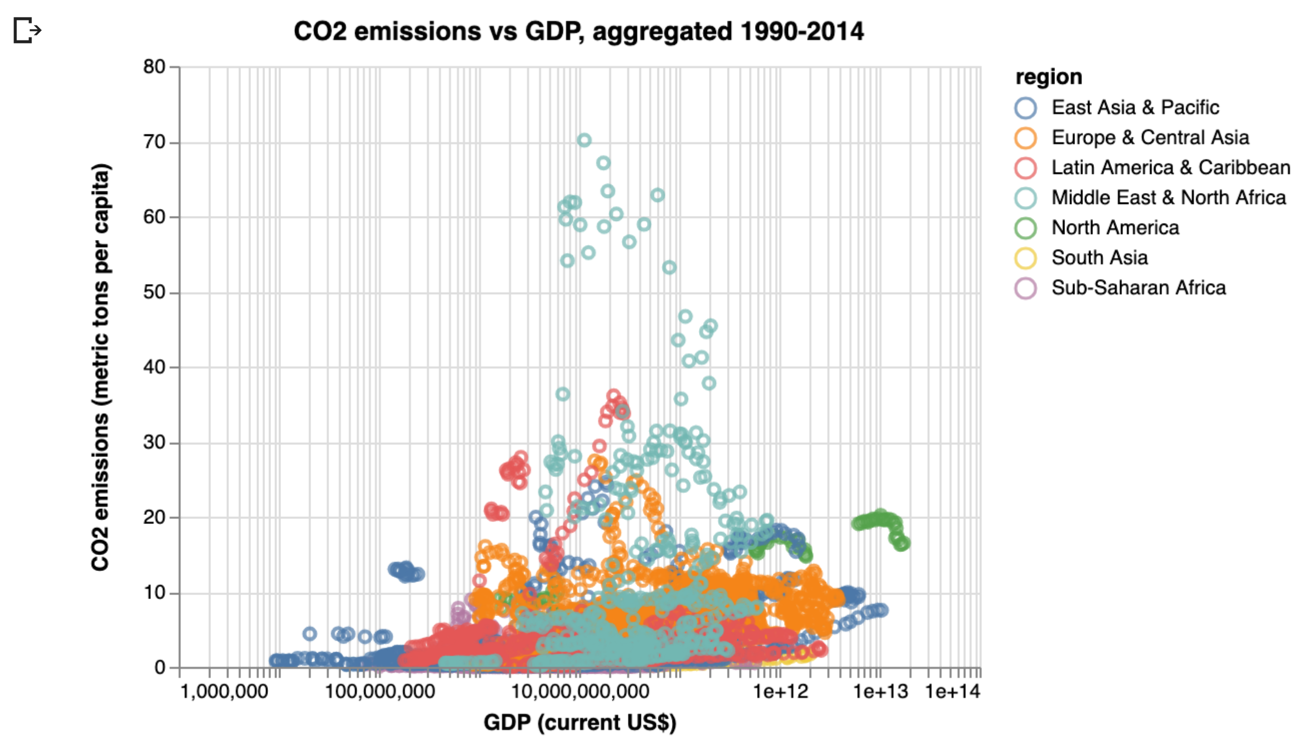
Export as SVG Export as PNG View Source Open in Vega Editor

We next look at the relationship between emissions and the nummber of researchers in R&D in each population). The graph seems to suggest that among the countries that emit around 5-10 mt per capi range of the number of researchers in R&D. The highest emitting countries (Kuwait and Qatar) have v R&D. However, the next highest emitters (US, Australia, Canada, Luxemborg), have a very similar numl R&D-- between 3,000 and 5,000 per million people.

```
%%bigquery --project $project_id p11

SELECT  gdp.value as gdp, gdp.country_name, co2.value as co2_pc, summary.region, gdp.
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` gdp,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
```

```
WHERE
   gdp.indicator_code= "NY.GDP.MKTP.CD" AND
   gdp.country_code= co2.country_code AND
   gdp.year= co2.year AND
   co2.indicator_code = "EN.ATM.CO2E.PC" AND
   summary.country_code= gdp.country_code AND
   summary.country_code= co2.country_code AND
   summary.region != "" AND
   co2.year >1990 AND gdp.year > 1990
```

```
alt.Chart(p11, title='CO2 emissions vs GDP, aggregated 1990-2014').mark_point().encod
    tooltip=['country_name',"year"],
    x=alt.X("gdp", axis=alt.Axis(title='GDP (current US$)'),scale=alt.Scale(type='log
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```



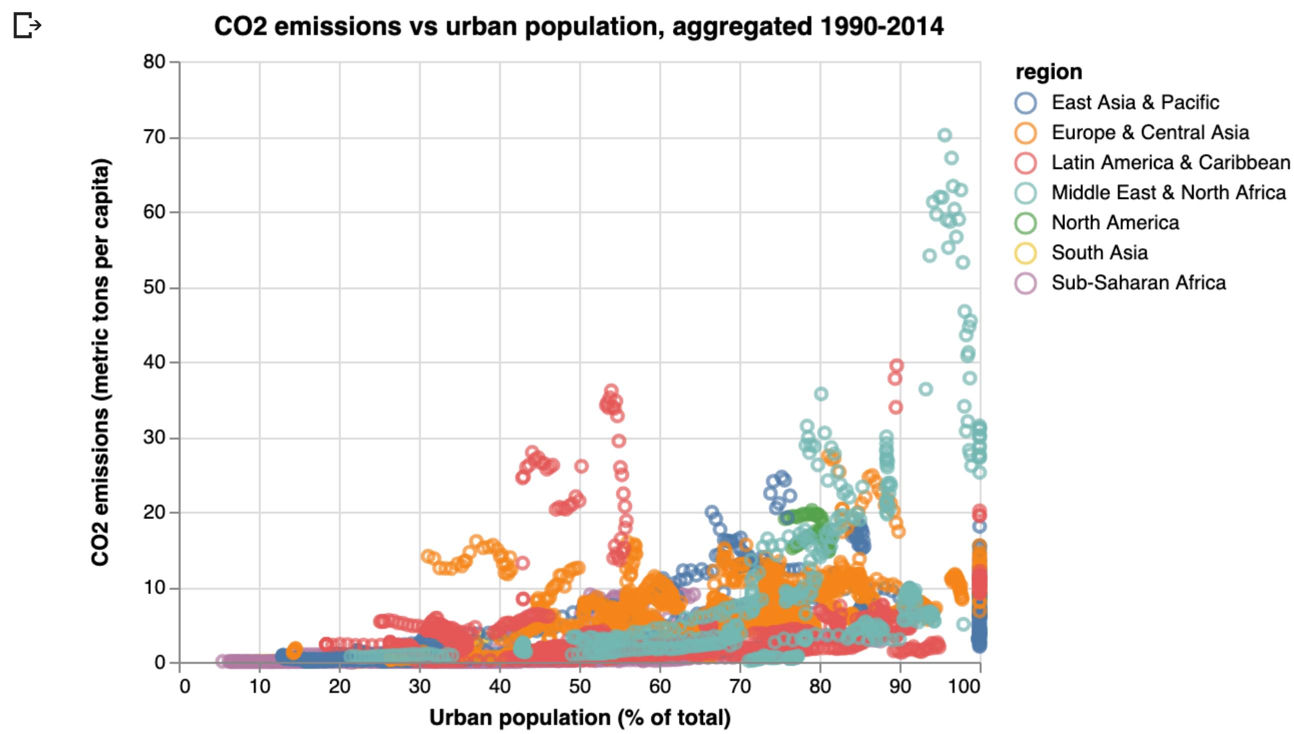
Export as SVG Export as PNG View Source Open in Vega Editor

Next, we look at the relationship between per capita emissions and log GDP. It is hard to get a clear re
graph. The US has the largest GDP, but countries with a GDP nearly 100x smaller (in Middle East, Cari
higher emissions. There do exist countries with low GDP that have emissions similar to that of the US

```
%%bigquery --project $project_id p12

SELECT  urban.value as urban, urban.country_name, co2.value as co2_pc, summary.region
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` urban,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
   urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
```

```
    urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
    urban.country_code= co2.country_code AND
    urban.year= co2.year AND
    co2.indicator_code = "EN.ATM.CO2E.PC" AND
    summary.country_code= urban.country_code AND
    summary.country_code= co2.country_code AND
    summary.region != "" AND
    co2.year >=1990 AND urban.year > 1990
```

```
alt.Chart(p12, title='CO2 emissions vs urban population, aggregated 1990-2014').mark_
    tooltip=['country_name','year'],
    x=alt.X("urban", axis=alt.Axis(title='Urban population (% of total)')),
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```



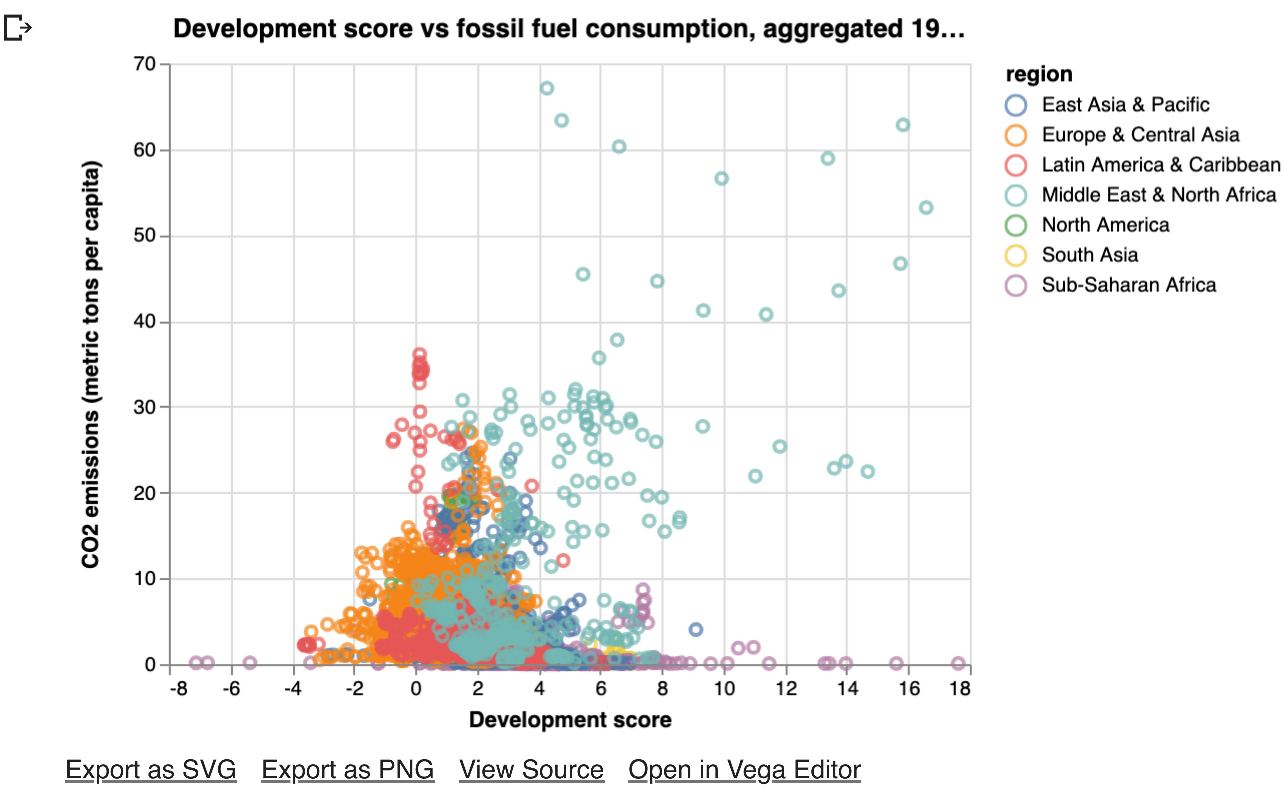
Export as SVG Export as PNG View Source Open in Vega Editor

We now look at the role of urban population on emisisons. Countries with larger emissions tend to ha
of people living in urban areas. The countries that emit the least tend to have a small urban populatio
suggest a positive relationship between urban population percentage and CO2 emissions.

```
%%bigquery --project $project_id p13

SELECT  urban.value as dev_score, gdp.country_name, co2.value as co2_pc, summary.regi
FROM
  `bigquery-public-data.world_bank_wdi.indicators_data` gdp,
  `bigquery-public-data.world_bank_wdi.indicators_data` urban,
  `bigquery-public-data.world_bank_wdi.indicators_data` co2,
  `bigquery-public-data.world_bank_health_population.country_summary` summary
WHERE
```

```
    gdp.indicator_code= "NY.GDP.MKTP.KD.ZG" AND
    gdp.country_code= co2.country_code AND
    gdp.year= co2.year AND
    urban.indicator_code= "SP.URB.GROW" AND
    urban.country_code= co2.country_code AND
    urban.year= co2.year AND
    co2.indicator_code = "EN.ATM.CO2E.PC" AND
    summary.country_code= gdp.country_code AND
    summary.country_code= co2.country_code AND
    summary.country_code= urban.country_code AND
    summary.region != "" AND
    co2.year >=1990 AND gdp.year >= 1990 AND urban.year >= 1990
```

```
lt.Chart(p13, title='Development score vs fossil fuel consumption, aggregated 1990-20
    tooltip=['country_name',"year"],
    x=alt.X("dev_score", axis=alt.Axis(title='Development score')),
    y=alt.Y("co2_pc", axis=alt.Axis(title='CO2 emissions (metric tons per capita)')),
  color='region')
```



Above, we develop the concept of an "development score", which is the product of a country's GDP gr urban population growth (annual %). It can be thought of as a measure of how quickly a country is ad emitters tend to have higher scores. However, countries like Rwanda have large scores but low emiss there is quite a bit of variability of emissions a country can make.

ML Predictions

Now that we have a better grasp of the relationship between several features and emissions, we are r
to predict emissions given certain features. We generate two models below.

```
# Run this cell to create a dataset to store your model, or create in the UI

model_dataset_name = 'co2_predict'

dataset = bigquery.Dataset(client.dataset(model_dataset_name))
dataset.location = 'US'
client.create_dataset(dataset)
```

Model 1

In our first model, we use the following features to predict per capita CO2 emissions:

- percentage of population living in urban area
- GDP
- percentage of energy coming from renewable resource
- percentage of population with access to electricity
- population
- development score
- percentage of energy consumption from fossil fuels
- country name
- year

We used a smaller subset of these original features in the final version of our model, after we did feat
dev set.

Each a sample consists of the feature information for a country in a given year. We use 80% of the da
reserve 10% for dev and 10% for test.

```
%%bigquery --project $project_id

# YOUR QUERY HERE

CREATE OR REPLACE MODEL `co2_predict.co2_model_3`
OPTIONS (model_type='linear_reg') AS
        -- TODO: write SQL to return the features and ground-truth values for the mod

SELECT urban.value as urban,
       renew.value as renew_elec,
       access.value as access_elec,
       co2.country_name as country,
       co2.year as year,
```

```
        ff.value as fossil_fuel,
        co2.value as label


    FROM `bigquery-public-data.world_bank_wdi.indicators_data` urban,
         `bigquery-public-data.world_bank_wdi.indicators_data` co2,
         `bigquery-public-data.world_bank_wdi.indicators_data` renew,
         `bigquery-public-data.world_bank_wdi.indicators_data` access,
         `bigquery-public-data.world_bank_wdi.indicators_data` ff,
         `bigquery-public-data.world_bank_health_population.country_summary` summary

    WHERE MOD(ABS(FARM_FINGERPRINT(co2.country_code)), 10) < 8 AND
                #urban constraints
                urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
                urban.country_code= co2.country_code AND
                urban.year= co2.year AND
                co2.indicator_code = "EN.ATM.CO2E.PC" AND
                #gdp constraints
                #renewable energy contraints
                renew.indicator_code= "EG.FEC.RNEW.ZS" AND
                renew.country_code= co2.country_code AND
                renew.year= co2.year AND
                #electricity access contraints
                access.indicator_code= "EG.ELC.ACCS.ZS" AND
                access.country_code= co2.country_code AND
                access.year= co2.year AND
                #fossil fuel growth constraints
                ff.indicator_code= "EG.USE.COMM.FO.ZS" AND
                ff.country_code= co2.country_code AND
                ff.year= co2.year AND
                #get rid of all the non countries
                summary.country_code= co2.country_code AND
                summary.country_code= urban.country_code AND
                summary.country_code= renew.country_code AND
                summary.country_code= access.country_code AND
                summary.country_code= ff.country_code AND
                summary.region != ""
```

We train the model below.

```
%%bigquery --project $project_id

# Run cell to view training stats

SELECT
  *
FROM
  ML.TRAINING_INFO(MODEL `co2_predict.co2_model_3`)
```

	training_run	iteration	loss	eval_loss	duration_ms	learning_rate
0	0	4	1.527898	1.960547	3038	0.4
1	0	3	1.698430	1.985395	3086	0.2
2	0	2	2.740310	2.876237	2712	0.8
3	0	1	8.225727	6.640097	2662	0.4
4	0	0	28.776927	26.360173	3016	0.2

We evaluate our model on the dev set.

```
%%bigquery --project $project_id

SELECT *
FROM
    ML.EVALUATE(MODEL `co2_predict.co2_model_3`, (
SELECT urban.value as urban,
       #gdp.value as gdp,
       renew.value as renew_elec,
       access.value as access_elec,
       #pop.value as population,
       co2.country_name as country,
       co2.year as year,
       #urban_growth.value * gdp_growth.value as dev_score,
       ff.value as fossil_fuel,
       co2.value as label


    FROM `bigquery-public-data.world_bank_wdi.indicators_data` urban,
         `bigquery-public-data.world_bank_wdi.indicators_data` co2,
         #`bigquery-public-data.world_bank_wdi.indicators_data` gdp,
         `bigquery-public-data.world_bank_wdi.indicators_data` renew,
         `bigquery-public-data.world_bank_wdi.indicators_data` access,
         #`bigquery-public-data.world_bank_wdi.indicators_data` pop,
         #`bigquery-public-data.world_bank_wdi.indicators_data` urban_growth,
         #`bigquery-public-data.world_bank_wdi.indicators_data` gdp_growth,
         `bigquery-public-data.world_bank_wdi.indicators_data` ff,
         `bigquery-public-data.world_bank_health_population.country_summary` summary

    WHERE MOD(ABS(FARM_FINGERPRINT(co2.country_code)), 10) = 8 AND
                #urban constraints
                urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
                urban.country_code= co2.country_code AND
                urban.year= co2.year AND
                co2.indicator_code = "EN.ATM.CO2E.PC" AND
                #gdp constraints
--                 gdp.indicator_code= "NY.GDP.MKTP.CD" AND
```

```
--              gdp.country_code= co2.country_code AND
--              gdp.year= co2.year AND
            #renewable energy contraints
            renew.indicator_code= "EG.FEC.RNEW.ZS" AND
            renew.country_code= co2.country_code AND
            renew.year= co2.year AND
            #electricity access contraints
            access.indicator_code= "EG.ELC.ACCS.ZS" AND
            access.country_code= co2.country_code AND
            access.year= co2.year AND
            #population constraints
--              pop.indicator_code= "SP.POP.TOTL" AND
--              pop.country_code= co2.country_code AND
--              pop.year= co2.year AND
            #urban growth constraints
--              urban_growth.indicator_code= "SP.URB.GROW" AND
--              urban_growth.country_code= co2.country_code AND
--              urban_growth.year= co2.year AND
--              #gdp growth constraints
--              gdp_growth.indicator_code= "NY.GDP.MKTP.KD.ZG" AND
--              gdp_growth.country_code= co2.country_code AND
--              gdp_growth.year= co2.year AND
            #fossil fuel growth constraints
            ff.indicator_code= "EG.USE.COMM.FO.ZS" AND
            ff.country_code= co2.country_code AND
            ff.year= co2.year AND
            #get rid of all the non countries
            summary.country_code= co2.country_code AND
            summary.country_code= urban.country_code AND
            #summary.country_code= gdp.country_code AND
            summary.country_code= renew.country_code AND
            summary.country_code= access.country_code AND
            #summary.country_code= pop.country_code AND
            #summary.country_code= urban_growth.country_code AND
            #summary.country_code= gdp_growth.country_code AND
            summary.country_code= ff.country_code AND
            summary.region != ""))
```

	mean_absolute_error	mean_squared_error	mean_squared_log_error	median_absolu
0	3.985463	55.566304	0.386584	

Let us examine the performance of our model. We perform linear regression and find we have an MSI which is not great given the scale of prediction we are making. Our r2 score is 0.16, which is also not variance is 0.33.

We tried several variations of this model. Our original model used number of people in R&D as a featu performance, since it decreased our dataset a lot. So, we decided to remove it, which improved perfo versions where we used log(population), or log(GDP), or excluded certain features, such as developm

and GDP. We evaluated the effect of those changes by looking at performance on our dev set, which i
Because our original dataset is roughly 3600 points, this does not leave a lot of data for training and t
our performance is so poor- with more data, the errors would likely be lower.

```
%%bigquery --project $project_id

SELECT *
FROM
    ML.EVALUATE(MODEL `co2_predict.co2_model_3`, (
SELECT urban.value as urban,
       renew.value as renew_elec,
       access.value as access_elec,
       co2.country_name as country,
       co2.year as year,
       ff.value as fossil_fuel,
       co2.value as label


    FROM `bigquery-public-data.world_bank_wdi.indicators_data` urban,
         `bigquery-public-data.world_bank_wdi.indicators_data` co2,
         `bigquery-public-data.world_bank_wdi.indicators_data` renew,
         `bigquery-public-data.world_bank_wdi.indicators_data` access,
         `bigquery-public-data.world_bank_wdi.indicators_data` ff,
         `bigquery-public-data.world_bank_health_population.country_summary` summary

    WHERE MOD(ABS(FARM_FINGERPRINT(co2.country_code)), 10) = 9 AND
                #urban constraints
                urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
                urban.country_code= co2.country_code AND
                urban.year= co2.year AND
                co2.indicator_code = "EN.ATM.CO2E.PC" AND
                #renewable energy contraints
                renew.indicator_code= "EG.FEC.RNEW.ZS" AND
                renew.country_code= co2.country_code AND
                renew.year= co2.year AND
                #electricity access contraints
                access.indicator_code= "EG.ELC.ACCS.ZS" AND
                access.country_code= co2.country_code AND
                access.year= co2.year AND
                #fossil fuel growth constraints
                ff.indicator_code= "EG.USE.COMM.FO.ZS" AND
                ff.country_code= co2.country_code AND
                ff.year= co2.year AND
                #get rid of all the non countries
                summary.country_code= co2.country_code AND
                summary.country_code= urban.country_code AND
                summary.country_code= renew.country_code AND
                summary.country_code= access.country_code AND
                summary.country_code= ff.country_code AND
```

```
        summary.region != ""))
```

	mean_absolute_error	mean_squared_error	mean_squared_log_error	median_absolu
0	2.889089	18.75221	0.329744	

```
%%bigquery --project $project_id

# YOUR QUERY HERE

SELECT
   country, year, predicted_label, label
FROM
  ML.PREDICT(MODEL `co2_predict.co2_model_3`, (

SELECT urban.value as urban,
       #gdp.value as gdp,
       renew.value as renew_elec,
       access.value as access_elec,
       #pop.value as population,
       co2.country_name as country,
       co2.year as year,
       #urban_growth.value * gdp_growth.value as dev_score,
       ff.value as fossil_fuel,
       co2.value as label


    FROM `bigquery-public-data.world_bank_wdi.indicators_data` urban,
         `bigquery-public-data.world_bank_wdi.indicators_data` co2,
         `bigquery-public-data.world_bank_wdi.indicators_data` renew,
         `bigquery-public-data.world_bank_wdi.indicators_data` access,
         `bigquery-public-data.world_bank_wdi.indicators_data` ff,
         `bigquery-public-data.world_bank_health_population.country_summary` summary

    WHERE MOD(ABS(FARM_FINGERPRINT(co2.country_code)), 10) = 8 AND
                #urban constraints
                urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
                urban.country_code= co2.country_code AND
                urban.year= co2.year AND
                co2.indicator_code = "EN.ATM.CO2E.PC" AND
                #renewable energy contraints
                renew.indicator_code= "EG.FEC.RNEW.ZS" AND
                renew.country_code= co2.country_code AND
                renew.year= co2.year AND
                #electricity access contraints
                access.indicator_code= "EG.ELC.ACCS.ZS" AND
                access.country_code= co2.country_code AND
                access.year= co2.year AND
                #fossil fuel growth constraints
                ff.indicator_code= "EG.USE.COMM.FO.ZS" AND
```

```
            ff.country_code= co2.country_code AND
            ff.year= co2.year AND
            #get rid of all the non countries
            summary.country_code= co2.country_code AND
            summary.country_code= urban.country_code AND
            summary.country_code= renew.country_code AND
            summary.country_code= access.country_code AND
            summary.country_code= ff.country_code AND
            summary.region != "" ))
LIMIT 20
```

	country	year	predicted_label	label
0	Chile	2009	4.680520	3.969556
1	Israel	2014	6.107909	7.863181
2	Israel	2004	6.213943	8.667993
3	Israel	2012	6.175866	9.547968
4	Jordan	2003	5.548557	3.237043
5	Jordan	1991	5.554685	2.610470
6	Kuwait	1996	7.063231	30.736244
7	Kuwait	2001	7.050211	27.326634
8	Mexico	1997	4.869809	3.801014
9	Mexico	1991	4.845240	3.812485
10	Ireland	1994	4.303759	9.100127
11	Paraguay	2010	1.209410	0.820810
12	Costa Rica	2000	2.828765	1.394704
13	Switzerland	1994	4.248081	5.908584
14	Macedonia, FYR	2011	3.329959	4.535127
15	Macedonia, FYR	1994	4.004704	5.282974
16	Trinidad and Tobago	1992	4.142594	15.477125
17	Trinidad and Tobago	2014	3.900080	34.163243
18	Trinidad and Tobago	2013	3.928790	34.520032
19	Trinidad and Tobago	1997	4.179669	14.590301

Finally, we evaluate our model on our test set, which is 10% of our unseen data. The model has a r2 o This is better than the model performed ont eh dev set, which is a bit strange, but could be due to the more outliers and harder countries to predict in the dev set, and easier examples in the test set.

We then look at some specific predictions our model makes. In many cases, the prediction is in the c
example, it correctly predicts small values for countries like Guatemala and Nigeria. It predicts larger
countries like Venezuela, Mexico, and Chile, as one would expect. However, the model has a tough tin
such as Trinidad and Tobago and Kuwait.

Model 2

We next build a second model. Here, we focus on time-dependent prediction. That is, we use the CO2
previous year as a feature for prediction in the current year. Instead of randomly splitting 80% of our c
did for our first model, we instead use data from before 2010 as our training set (3599 data points)an
data after 2010 (722 data points). This corresponts to saving just under 20% for test.

We use the following features for this model:

- percentage of population living in urban area
- GDP
- percentage of energy coming from renewable resource
- percentage of population with access to electricity
- population
- CO2 emission of previous year
- country name
- year

Train the model

```
%%bigquery --project $project_id

# YOUR QUERY HERE

CREATE OR REPLACE MODEL `co2_predict.co2_model_forecast`
OPTIONS (model_type='linear_reg') AS
        -- TODO: write SQL to return the features and ground-truth values for the mod

SELECT urban.value as urban,
       gdp.value as gdp,
       renew.value as renew_elec,
       access.value as access_elec,
       pop.value as population,
       co2.country_name as country,
       co2.year as year,
       co2_prev.value as prev_co2,
       co2.value as label
```

```
FROM `bigquery-public-data.world_bank_wdi.indicators_data` urban,
     `bigquery-public-data.world_bank_wdi.indicators_data` co2,
     `bigquery-public-data.world_bank_wdi.indicators_data` gdp,
     `bigquery-public-data.world_bank_wdi.indicators_data` renew,
     `bigquery-public-data.world_bank_wdi.indicators_data` access,
     `bigquery-public-data.world_bank_wdi.indicators_data` pop,
     `bigquery-public-data.world_bank_wdi.indicators_data` co2_prev,
     `bigquery-public-data.world_bank_health_population.country_summary` summary

WHERE
            #urban constraints
            urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
            urban.country_code= co2.country_code AND
            urban.year= co2.year AND
            co2.indicator_code = "EN.ATM.CO2E.PC" AND
            #gdp constraints
            gdp.indicator_code= "NY.GDP.MKTP.CD" AND
            gdp.country_code= co2.country_code AND
            gdp.year= co2.year AND
            #renewable energy contraints
            renew.indicator_code= "EG.FEC.RNEW.ZS" AND
            renew.country_code= co2.country_code AND
            renew.year= co2.year AND
            #electricity access contraints
            access.indicator_code= "EG.ELC.ACCS.ZS" AND
            access.country_code= co2.country_code AND
            access.year= co2.year AND
            #population constraints
            pop.indicator_code= "SP.POP.TOTL" AND
            pop.country_code= co2.country_code AND
            pop.year= co2.year AND
            #previous year's emissions constraints
            co2_prev.indicator_code= "EN.ATM.CO2E.PC" AND
            co2_prev.country_code= co2.country_code AND
            co2_prev.year - 1 = co2.year AND
            #get rid of all the non countries
            summary.country_code= co2.country_code AND
            summary.country_code= urban.country_code AND
            summary.country_code= gdp.country_code AND
            summary.country_code= renew.country_code AND
            summary.country_code= access.country_code AND
            summary.country_code= pop.country_code AND
            summary.region != "" AND
            #for training, use years before 2010
            co2.year < 2010


%%bigquery --project $project_id

# Run cell to view training stats

SELECT
```

```
SELECT
  *
FROM
  ML.TRAINING_INFO(MODEL `co2_predict.co2_model_forecast`)
```

	training_run	iteration	loss	eval_loss	duration_ms	learning_rate
0	0	7	0.867610	1.065446	2725	0.8
1	0	6	0.871430	1.070993	2166	0.4
2	0	5	0.875746	1.099054	1895	0.2
3	0	4	0.888727	1.144982	1868	0.4
4	0	3	0.937236	1.168910	2368	0.4
5	0	2	1.175821	1.763217	2432	0.4
6	0	1	2.732571	3.657797	2254	0.4
7	0	0	17.789453	23.957417	2661	0.2

Now, evaluate the model

```
%%bigquery --project $project_id

SELECT *
FROM
    ML.EVALUATE(MODEL `co2_predict.co2_model_forecast`, (
SELECT urban.value as urban,
       gdp.value as gdp,
       renew.value as renew_elec,
       access.value as access_elec,
       pop.value as population,
       co2.country_name as country,
       co2.year as year,
       co2_prev.value as prev_co2,
       co2.value as label


FROM `bigquery-public-data.world_bank_wdi.indicators_data` urban,
     `bigquery-public-data.world_bank_wdi.indicators_data` co2,
     `bigquery-public-data.world_bank_wdi.indicators_data` gdp,
     `bigquery-public-data.world_bank_wdi.indicators_data` renew,
     `bigquery-public-data.world_bank_wdi.indicators_data` access,
     `bigquery-public-data.world_bank_wdi.indicators_data` pop,
     `bigquery-public-data.world_bank_wdi.indicators_data` co2_prev,
     `bigquery-public-data.world_bank_health_population.country_summary` summary

WHERE
            #urban constraints
            urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
```

```
        urban.country_code= co2.country_code AND
        urban.year= co2.year AND
        co2.indicator_code = "EN.ATM.CO2E.PC" AND
        #gdp constraints
        gdp.indicator_code= "NY.GDP.MKTP.CD" AND
        gdp.country_code= co2.country_code AND
        gdp.year= co2.year AND
        #renewable energy contraints
        renew.indicator_code= "EG.FEC.RNEW.ZS" AND
        renew.country_code= co2.country_code AND
        renew.year= co2.year AND
        #electricity access contraints
        access.indicator_code= "EG.ELC.ACCS.ZS" AND
        access.country_code= co2.country_code AND
        access.year= co2.year AND
        #population constraints
        pop.indicator_code= "SP.POP.TOTL" AND
        pop.country_code= co2.country_code AND
        pop.year= co2.year AND
        #previous year's emissions constraints
        co2_prev.indicator_code= "EN.ATM.CO2E.PC" AND
        co2_prev.country_code= co2.country_code AND
        co2_prev.year - 1 = co2.year AND
        #get rid of all the non countries
        summary.country_code= co2.country_code AND
        summary.country_code= urban.country_code AND
        summary.country_code= gdp.country_code AND
        summary.country_code= renew.country_code AND
        summary.country_code= access.country_code AND
        summary.country_code= pop.country_code AND
        summary.region != "" AND
        #for testing, use years after 2010
        co2.year >= 2010 ))
```

	mean_absolute_error	mean_squared_error	mean_squared_log_error	median_absolu
0	0.580398	1.659237	0.022226	

This model does a much better job, since it gets the CO2 emission from the previous year as a featur an r2 score of 0.96, which is pretty good. The MSE is also much smaller than that of the first model.

Now, do prediction

```
%%bigquery --project $project_id

# YOUR QUERY HERE

SELECT
```

```
   country, year, predicted_label, label
FROM
  ML.PREDICT(MODEL `co2_predict.co2_model_forecast`, (

SELECT urban.value as urban,
       gdp.value as gdp,
       renew.value as renew_elec,
       access.value as access_elec,
       pop.value as population,
       co2.country_name as country,
       co2.year as year,
       co2_prev.value as prev_co2,
       co2.value as label



FROM `bigquery-public-data.world_bank_wdi.indicators_data` urban,
     `bigquery-public-data.world_bank_wdi.indicators_data` co2,
     `bigquery-public-data.world_bank_wdi.indicators_data` gdp,
     `bigquery-public-data.world_bank_wdi.indicators_data` renew,
     `bigquery-public-data.world_bank_wdi.indicators_data` access,
     `bigquery-public-data.world_bank_wdi.indicators_data` pop,
     `bigquery-public-data.world_bank_wdi.indicators_data` co2_prev,
     `bigquery-public-data.world_bank_health_population.country_summary` summary

WHERE
            #urban constraints
            urban.indicator_code= "SP.URB.TOTL.IN.ZS" AND
            urban.country_code= co2.country_code AND
            urban.year= co2.year AND
            co2.indicator_code = "EN.ATM.CO2E.PC" AND
            #gdp constraints
            gdp.indicator_code= "NY.GDP.MKTP.CD" AND
            gdp.country_code= co2.country_code AND
            gdp.year= co2.year AND
            #renewable energy contraints
            renew.indicator_code= "EG.FEC.RNEW.ZS" AND
            renew.country_code= co2.country_code AND
            renew.year= co2.year AND
            #electricity access contraints
            access.indicator_code= "EG.ELC.ACCS.ZS" AND
            access.country_code= co2.country_code AND
            access.year= co2.year AND
            #population constraints
            pop.indicator_code= "SP.POP.TOTL" AND
            pop.country_code= co2.country_code AND
            pop.year= co2.year AND
            #previous year's emissions constraints
            co2_prev.indicator_code= "EN.ATM.CO2E.PC" AND
            co2_prev.country_code= co2.country_code AND
            co2_prev.year - 1 = co2.year AND
            #get rid of all the non countries
            summary.country_code= co2.country_code AND
```

```
        summary.country_code= urban.country_code AND
        summary.country_code= gdp.country_code AND
        summary.country_code= renew.country_code AND
        summary.country_code= access.country_code AND
        summary.country_code= pop.country_code AND
        summary.region != "" AND
        #for testing, use years after 2010
        co2.year >= 2010 ))
LIMIT 20
```

	country	year	predicted_label	label
0	Cameroon	2010	0.578512	0.339515
1	Sierra Leone	2010	0.244201	0.112416
2	Cambodia	2010	0.542841	0.350331
3	Macedonia, FYR	2011	4.812412	4.535127
4	Angola	2011	1.644593	1.252789
5	Ecuador	2011	2.228186	2.543911
6	Samoa	2011	0.961219	1.074708
7	Greenland	2011	9.762608	12.440341
8	Haiti	2012	0.461023	0.224884
9	Ghana	2012	1.360563	0.461563
10	Nigeria	2012	0.978997	0.588790
11	Grenada	2012	2.161370	2.572577
12	Sao Tome and Principe	2012	0.849758	0.621563
13	Greenland	2013	9.304293	9.803251
14	Bahamas, The	2013	5.933691	7.426540
15	Iran, Islamic Rep.	2013	7.014397	8.003809
16	Niger	2013	0.245967	0.105275
17	Qatar	2013	50.931245	37.780085
18	Trinidad and Tobago	2013	24.704186	34.520032
19	Colombia	2013	1.883549	1.893103

We look at some specific predictions our model makes. In most cases, the prediction is pretty accura right range. We see the model again has trouble with Qatar and Trinidad and Tobago, outliers in many that are one of the largest emitters. But for the most part, this model is able to predict in the rough ba

Conclusions

In this project, we used BigQuery to make predictions about worldwide CO2 emissions per capita. Firs visualizations to get a better sense of worldwide CO2 emissions over the past 50 years. Then, we cre to understand the relationship between emissions and several features. Some graphs were more clea others. For example, there was a more clear relationship with certain features like levels of urbanizati electricity. Other variables, like GDP and population, were harder to make sense of.

After we gained greater insights between per capita CO2 emissions and several features, we generate models to perform prediction to answer our questions. Our first model used a set of 9 feautres to pre some major trouble with some countries that were unique outliers in the visualizations we had create understandable. We then created a model to predict CO2 emissions given the previous year's emissic model on data from before 2010 and tested on data after 2010. This model did a much better job, sin previous year's emission, which is probably a very helpful feature to have for the model. The model st the outlier countries, but its performance was stronger than that of the first model.

It would have been intersting to do include features related to education in the model. We had origina in earlier iterations, but it really decreased the size of our dataset. Perhaps there are other datasets re information, and if we had more time, we could integrate that information into our model.