select race,

count (voterbase\_id) as population

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by race

SELECT gender,

race,

count (voterbase\_id) as population

FROM targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by gender, race

order by gender

select race,

voter\_type,

count (voterbase\_id) as population

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by race, voter\_type

select count (voterbase\_id)

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

order by 1

select age\_bucket,

count (\*)

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by age\_bucket

select count (voterbase\_id) as population,

county

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by county

select count (voterbase\_id) as population,

county,

race

from targets\_2022.base\_file

where vf\_source\_state = 'AZ' and county = 'YUMA'

group by county, race

select

county,

avg (part\_score) as part\_score\_avg,

avg (turnout\_score) as turnout\_score\_avg

from targets\_2022.base\_file

where county = 'PIMA' and vf\_source\_state = 'AZ'

group by 1

select

county,

avg (part\_score) as part\_score\_avg,

avg (turnout\_score) as turnout\_score\_avg

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by 1

select

race,

avg (part\_score) as part\_score\_avg,

avg (turnout\_score) as turnout\_score\_avg

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by 1

select

count (voterbase\_id),

race,

zip,

vf\_source\_state

from targets\_2022.base\_file

where vf\_source\_state = 'AZ' and race = 'Hispanic'

group by race, zip, vf\_source\_state

select

count (voterbase\_id),

race,

zip,

vf\_source\_state

from targets\_2022.base\_file

where vf\_source\_state = 'AZ' and race = 'AfAm'

group by race, zip, vf\_source\_state

select

ceil (persuasion\_score \* 10)/10 pers\_bucket,

count (\*)

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by 1

select

ceil (pers\_rank \* 10)/10 gotvbucket,

count (\*)

from targets\_2022.base\_file

where vf\_source\_state = 'AZ' and gotv\_score > 0

group by 1

select

avg (part\_score) as part\_score\_avg,

avg (gotv\_score) as gotv\_score\_avg,

avg (turnout\_score) as turnout\_score\_avg,

race,

income\_bucket

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by race, income\_bucket

select

case when pers\_rank <= 500000 then 'top 500k' -- not working, might be because of NULLs ???

when race = 'White' and pers\_percentile\_rank >= 0.80 then 'top 25% of targets, white'

when race != 'White' and pers\_percentile\_rank >= 0.70 then 'top 30% of targets, non-white'

else 'non-target'

end as audience,

count(\*) as pop,

count(case when race = 'White' then voterbase\_id end) as pop\_white,

count(case when race != 'White' then voterbase\_id end) as pop\_nonwhite,

avg(part\_score) as avg\_part\_score,

avg(turnout\_score) as avg\_turnout\_score,

avg(persuasion\_score) as avg\_persuasion\_score

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by 1

select count (voterbase\_id),

edu\_college,

voter\_type

from targets\_2022.base\_file

where vf\_source\_state = 'AZ' and edu\_college = 'college'

group by edu\_college, voter\_type

--UPDATED CODE 10/21/22 NEEDED TO KEEP GROUPING VARIABLE AND

-- AGGREGATE FUNTIONS TOGETHER AKA RACE/AGE\_BUCKET THEN turnout\_score

-- PART\_SCORE AND CASE WHEN STATEMENTS.

--ANY GROUPING VARIABLES LIKE AGE\_BUCKET, RACE, INCOME ETC NEED TO BE LISTED

--IN THE GROUP BY STATEMENT AT THE END OF THE CODE

--COUNT \* IS PULLING IN INFORMATION FROM ALL COLUMNS.

--TURNOUT SCORE NEEDS TO BE = 0-1

--ADD INTO GOOGLE SHEETS TO START MANIPULATING THE DATA

select

race,

case

when race = 'White'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .15

then 'white persuasion'

when race = 'AfAm'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .15

then 'AfAm persuasion'

when race = 'Hispanic'

and part\_score >=.3

and part\_score <.7

and turnout\_score >.15

then 'Hispanic persuasion'

when race = 'Native'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .15

then 'Native Persuasion'

when race = 'Asian'

and part\_score >= .3

and part\_score <.7

and turnout\_score > .15

then 'Asian persuasion'

else 'non\_target'

end as audience,

count(voterbase\_id),

avg(part\_score) as avg\_part\_score,

avg(turnout\_score) as avg\_turnout\_score

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by race, audience

UPDATED CODE AS OF OCTOBER 24TH 2022

GROUPING BY RACE, VOTER\_TYPE, GENDER, AND AUDIENCE.

select

race,

gender,

voter\_type,

case

when race = 'White'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .15

then 'white persuasion'

when race = 'AfAm'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .15

then 'AfAm persuasion'

when race = 'Hispanic'

and part\_score >=.3

and part\_score <.7

and turnout\_score >.15

then 'Hispanic persuasion'

when race = 'Native'

and part\_score >=.3

and part\_score <.7

then 'Native Persuasion'

when race = 'Asian'

and part\_score >= .3

and part\_score <.7

and turnout\_score > .15

then 'Asian persuasion'

else 'non\_target'

end as audience,

count(voterbase\_id),

avg(part\_score) as avg\_part\_score,

avg(turnout\_score) as avg\_turnout\_score

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by race, gender, voter\_type, audience

SQL Practice Problems

Instructions

Living Document - place for Anisha to put practice problems related to different commands learned through mode training. You do not need to complete all the problems! Do as many as is necessary to understand the concept.

Make a duplicate of this document and use it to post answers and write notes! Make this into an easy reference sheet for yourself as you move forward through the fellowship.

Note: before you try to answer the questions, make sure you understand the data you are querying. Look at a few rows, and make sure you know what each column represents, and what each row represents. Ask questions if you are confused!

How to post answers:

\* When the query results in a table, paste the table as the answer or paste a link to the query on github as the answer.

\* When the query results in a single number, paste the number as the answer.

How to save answers:

\* Please save a single file full of queries for all the below problems to github.

\* Link to github file ← please add this for Anisha to easily review!

Where to Query:

\* Table to query (unless otherwise specified): targets\_2022.base\_file

Basic SQL: Manipulate Data

Query table: scratch.sql\_practice

Hint: these problems may require using Order By, Arithmetic, and Where statements that leverage AND/OR

Example Problem:

1. What is the average partisanship score for 18-34 white voters in WI?

Answer: 0.538

Query:

select

avg\_part\_score,

\*

from scratch.sql\_practice

where vf\_source\_state = 'WI' and race = 'White' and age\_bucket = '18-34'

;

Practice Problems:

2. What are the average partisanship scores by age for all white voters in WI? Your query will return a table of data—order the rows by age to sort it youngest to oldest

avg\_part\_score age\_bucket vf\_source\_state race

0.538101145662515 18-34 WI White

0.516410841571253 35-49 WI White

0.46268441934293 50-64 WI White

0.488928073112574 65+ WI White

Query:

select

avg\_part\_score,

age\_bucket,

vf\_source\_state,

race

FROM scratch.sql\_practice

where (vf\_source\_state = 'WI' and race = 'White')

order by age\_bucket

3. How many Hispanic people are college educated in AZ?

Sum = 93778

SELECT SUM (pop\_college)

FROM scratch.sql\_practice

where race = 'Hispanic' and vf\_source\_state='AZ'

;

4. How many voters 18-34 are college educated in NH?

Sum = 7600

select sum (pop\_college)

from scratch.sql\_practice

where age\_bucket = '18-34' and vf\_source\_state = 'NH'

;

5. Among Black voters in MI and WI, what age bucket has the highest turnout score?

Highest Age Turnout score : 50-64

select avg\_turnout\_score,

\*

from scratch.sql\_practice

where (vf\_source\_state = 'MI' and race='AfAm') or (vf\_source\_state = 'WI' and race ='AfAm')

order by avg\_turnout\_score desc

6. How many college-educated people in NV are aged 18-34 or Hispanic?

A: 42192

select sum (pop\_college)

from scratch.sql\_practice

where (vf\_source\_state = 'NV') and (age\_bucket = '18-34'or race = 'Hispanic')

7. How many women in PA are 35-49 or Asian in PA? (Hint: This query will require using the AND function together with the OR function, and then adding up the resulting rows)

A: 1293359

select sum (pop\_female)

from scratch.sql\_practice

where (race = 'Asian' or age\_bucket = '35-49')

and vf\_source\_state = 'PA'

8. Create a column that calculates the % of people who are college educated

select (pop\_college::decimal/pop) AS Per\_coll

from scratch.sql\_practice

9. Create a column that calculates the sum of women and parents

select pop\_college + pop\_parent AS pop\_collpar

from scratch.sql\_practice

Intermediate SQL: Aggregate Functions

Support score=Part\_score

1. Count / Sum / Min / Max – no Group By

Result will be a single line

1. How many people are in the table?

select

count (voterbase\_id) as number\_of\_people

from targets\_2022.base\_file

43164037

2. How many women are in the table?

select gender,

count (voterbase\_id)

from targets\_2022.base\_file

group by 1

Female: 22357905

3. What is the minimum and maximum gotv\_score in the table?

select

min(gotv\_score) as min\_score,

max(gotv\_score) as max\_score

from targets\_2022.base\_file

MIN: -0.168698420810276

MAX: 0.168749040229618

4. What is the sum of turnout in the state of AZ?

select

vf\_source\_state,

sum(turnout\_score)

from targets\_2022.base\_file

group by vf\_source\_state

2591138.27974429

5. What is the average support score in NV?

select

vf\_source\_state,

avg (part\_score)

from targets\_2022.base\_file

group by 1

NV: 0.503069070629719

2. Count / Sum / Min / Max - with Group By

Result will be a table

1. How many people are there in each state?

select vf\_source\_state,

count (voterbase\_id) as number\_of\_people

from targets\_2022.base\_file

group by 1

NV 2613318

WI 5076912

MI 9206261

PA 10543765

AZ 5684660

GA 8915066

NH 1124055

2. How many women are there, by race?

select count (voterbase\_id),

race,

gender

from targets\_2022.base\_file

where gender = 'Female'

group by gender, race

count race gender

17194340 White Female

1364326 Hispanic Female

3097092 AfAm Female

623928 Asian Female

78219 Native Female

3. How many people are there by age and education?

select count (voterbase\_id),

age\_bucket,

edu\_college

from targets\_2022.base\_file

group by age\_bucket, edu\_college

count age\_bucket edu\_college

8693493 50-64 noncollege

8479393 35-49 noncollege

8967257 65+ noncollege

3255001 65+ college

420420 18-34 college

9452199 18-34 noncollege

2527721 50-64 college

1368553 35-49 college

4. What is the average support and turnout, by state?

select vf\_source\_state,

avg (part\_score) as part\_score\_avg,

avg (turnout\_score) as turnout\_score\_avg

from targets\_2022.base\_file

group by vf\_source\_state

vf\_source\_state part\_score\_avg turnout\_score\_avg

NV 0.503069070629719 0.391248780102607

WI 0.472362055995254 0.497718188488702

PA 0.46506544868448 0.613672246697333

MI 0.490640919327977 0.460100772757303

AZ 0.482996036443998 0.567034219103437

GA 0.53496228613047 0.520803251479176

NH 0.453490321479449 0.561661443998332

5. What is the sum of turnout, by state?

select vf\_source\_state,

sum (turnout\_score)

from targets\_2022.base\_file

group by vf\_source\_state

vf\_source\_state sum

NV 941392.688526825

WI 2491686.74957591

PA 5250878.60112652

MI 4134010.96443364

AZ 2591138.27974429

GA 4642995.35995145

NH 618133.1322237

3. Count / Sum / Min / Max - with Group By and Having

These are challenging ! Ask for clarification if needed

1. Which Congressional Districts cover more than 1 county?

1. Hint: Count how many counties there are, grouped by congressional district, and use having to filter for only congressional districts have more than 1

select count (Distinct county),

cd

from targets\_2022.base\_file

group by cd

having count (distinct county) > 1

count cd

12 PA-09

6 PA-16

187 NULL

11 MI-09

4 MI-13

6 MI-03

17 GA-11

37 GA-10

27 GA-03

8 GA-07

2. Which zip codes cover more than 1 county?

select count (distinct county),

zip

from targets\_2022.base\_file

group by zip

having count (distinct county) > 1

count zip

3 18419

2 15108

2 15146

2 19150

2 15321

2 19041

2 16441

2 15057

2 16159

2 18640

3. Which states have more than 5% Hispanic voters?

SELECT COUNT(\*) AS race\_pop, race, vf\_source\_state

FROM targets\_2022.base\_file

GROUP BY race, vf\_source\_state

Intermediate SQL: Case When

1. Case when

Result will be building a column.

1. Code anyone who is a woman and is college-educated as 1, code anyone who is not 0

SELECT gender,

edu\_college,

CASE WHEN (gender = 'Female' AND edu\_college = 'college')

THEN 1 ELSE 0 END

FROM targets\_2022.base\_file

2. Code anyone who is not white and 18-34 as ‘young POC’, code anyone who is not as ‘non-target’

select race,

age\_bucket,

case when (race = 'White' AND age\_bucket = '18-34')

then 'young POC' else 'non-target' end

from targets\_2022.base\_file

3. Code anyone who is Black or 18-34 as ‘young or Black’, code anyone who is not as ‘non-target’

select race,

age\_bucket,

case when (race = 'black' and age\_bucket = '18-34')

then 'young or black' else 'non-target' end

from targets\_2022.base\_file

4. Three levels: 1) anyone who is white & non-college educated with a support score of 60+, 2) anyone else who has a support score of 60+, 3) anyone else

2. Case when + Aggregate Function

Result will be a single number

1. Count how many people were coded 1, grouped by state, in the section above, Part A Question #1

2. Count how many people were coded ‘young POC’, grouped by state, in the section above, Part A Question #2

3. Count how many people were coded ‘young or Black’, grouped by state, in the section above, Part A Question #3

4. Count how many people were coded at each level, across MI, WI, PA, in the section above, Part A Question #4

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Intermediate SQL: Joins

Joins are tricky!! Making sure you understand the result of the join and that it matches the desired result can be a challenge. Sometimes code works and then it is actually very wrong!

All newbie SQL users have a venn diagram for joins to reference handy. Here is one that I like:

1. Inner Join:

1. How many targets in our mobilization universe in AZ are also in our persuasion universe in AZ?

1. Step 1: Join table targets\_2022.mob\_base\_file with targets\_2022.az\_smp\_accountability\_1p\_20220428 on voterbase\_id

2. Step 2: Filter the result where target\_type with mob\_base\_file is ‘mobilization’ and latino\_newsboost\_bucket within az\_smp\_accountability\_1p\_20220428 is not ‘non-target’

3. Step 3: Count how many records have been produced using an aggregate function with the select statement

2. What is the average support score by age\_bucket in MI?

1. Step 1: Join table scores\_2022.whitmer\_support\_20220428 with targets\_2022.base\_file on voterbase\_id

2. Step 2: Group by age\_bucket from base\_file

3. Step 3: Calculate average of gov\_dv\_1 from whitmer\_support\_20220428 using an aggregate function with the select statement

2. Left Join:

1. Join together data and create a column, “support\_score.” When a support score is available, use that score. When it is not available, use partisanship score.

1. Step 1: Left join table targets\_2022.base\_file with scores\_2022.br\_az\_senate\_20220222 on voterbase\_id. The latter table has support and turnout scores for ONLY AZ.

2. Step 2: Use CASE WHEN to create a new column “support\_score” that uses support from br\_az\_senate\_20220222 when it is not null and “part\_score” otherwise

3. Step 3: Group by state

4. Step 4: Calculate an average of “support\_score” for every state using an aggregate function within the select statement

UPDATED OCTOBER 24, 2022 KG SQL ASSESMENT PROBLEMS

--Using the table ts.basic\_noncommercial\_client:

--Write a SQL query to select all individuals (voterbase\_id) for a specific state. Choose from: AZ, GA, MI, NH, PA, WI

--Select columns for: race5way, gender, age, marriage, parent, partisanship, voter status, registration date

--Rename the columns (shorten the names) as you select them

select voterbase\_id,

race5way\_noncommercial as race,

gender\_noncommercial as gender,

coalesced\_noncommercial\_age as age,

marriage\_noncommercial as marriage\_status,

parent\_noncommercial as parental\_status,

partisanship\_dem\_pr\_noncommercial as part\_dem\_pr,

voter\_status,

registration\_date

from ts.basic\_noncommercial\_client

where vf\_source\_state = 'AZ'

order by 1

--Using ts.modeling\_noncommercial\_client:

--Join the following columns to the ones you selected in the previous step: ts\_midterm\_general\_turnout, education\_collegegrad, education\_highschool, ideology

--What type of join do you want to use here?

---KG AZ PP & PAC

--white college and non-college nbv. older 35+ hispanics and young 18-34 hispanics

--partisonship 30-70; Turnout 35+(W College\_edu) 40+(W Non-college\_nbv),

--20+(POC) 20+(H young\_18-34) 20+(H old\_35+) 20+(POC)

--USING GROUPING VARIABLES

select

case

when race = 'White'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .35

and edu\_college = 'college'

then 'white college persuasion'

when race = 'White'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .40

and edu\_college = 'noncollege'

and voter\_type is not null -- only NBV coalition

then 'white non-college persuasion, NBV coalition'

when race = 'Hispanic'

and part\_score >=.3

and part\_score <.7

and turnout\_score >.20

and age\_bucket = '18-34'

then 'Hispanic persuasion, young'

when race = 'Hispanic'

and part\_score >=.3

and part\_score <.7

and turnout\_score >.20

and age\_bucket != '18-34'

then 'Hispanic persuasion, 35+'

when race in ('AfAm','Native','Asian')

and part\_score >=.3

and part\_score <.7

and turnout\_score > .20

then 'Other POC persuasion'

else 'non\_target'

end as audience,

count(voterbase\_id) as pop,

avg(part\_score) as avg\_part\_score,

avg(turnout\_score) as avg\_turnout\_score,

avg(persuasion\_score) as persuasion\_score

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by audience

--NO GROUPING VARIABLES ONLY SCORES TO CALCULATE

select

case

when race = 'White'

and part\_score >=.3

and part\_score <.7

and turnout\_score > .35

then 'white persuasion'

when race = 'Hispanic'

and part\_score >=.3

and part\_score <.7

and turnout\_score >.20

then 'Hispanic persuasion'

when race in ('AfAm','Native','Asian')

and part\_score >=.3

and part\_score <.7

and turnout\_score > .20

then 'Other POC persuasion'

else 'non\_target'

end as audience,

count(voterbase\_id) as pop,

avg(part\_score) as avg\_part\_score,

avg(turnout\_score) as avg\_turnout\_score,

avg(persuasion\_score) as persuasion\_score

from targets\_2022.base\_file

where vf\_source\_state = 'AZ'

group by audience