Final Report

Your final submission to Brightspace should be a link to your GitHub repository.
The repository should include:
• A readme file to briefly introduce the theme and purpose of your project, and instructions for using your code.
• Your final project report (a pdf file). More instructions below.
• Your code (organized in folders with meaningful names, as you see fit).

Your final report should contain the following sections:

Background

Our project, centered on Option C: YouTube Activity Analysis for Researchers, Journalists, etc., was motivated by the desire to approach YouTube with a researcher's mindset. Our primary objective involved looking into the creators, videos, and categories to extract valuable insights that could answer specific questions and inform strategic content decisions.

Focusing on systematic information gathering regarding creators, individual videos, and content categorization, our analysis aimed to uncover patterns of content engagement, geographical trends, and identify preferences within distinct video categories.

Addressing key questions about creator insights, user engagement, and category roles, our data analysis encompassed content engagement metrics, geographical trends, and category preferences. The dual purpose of this data was to refine content strategies and enhance our understanding of target demographics. In summary, our research provides strategic insights that not only contribute to refined content strategies but also offer a deeper understanding of demographics, ultimately assisting in identifying popular content genres on the platform.

Database Description

Creator Table:

- Username (VARCHAR2(255)) - Primary Key: Unique identifier for the user (YouTube username).
- NumOfSubscribers (INTEGER) - Default 0: Number of subscribers the user has.
- JoinDate (DATE) - NOT NULL: Date when the creator joined YouTube.
- TotalViews (INTEGER) - Default 0: Total views across all videos by the creator.
- Country (VARCHAR2(50)) - NOT NULL: Country of residence of the creator.
- TotalVideos (INTEGER) - Default 0: Total number of videos uploaded by the creator.

Video Table:

- VideoID (VARCHAR2(150)) - Primary Key: Unique identifier for each video.
- CategoryID (INTEGER) - Foreign Key: References the Category table.

- Username (NVARCHAR2(50)) - NOT NULL - Foreign Key: Username of the video uploader.
- PublishedOn (DATE) - NOT NULL: Date when the video was published.
- NumOfComments (INTEGER) - Default 0: Number of comments on the video.
- NumOfLikes (INTEGER) - Default 0: Number of likes on the video.

Category Table:

- CategoryID (INTEGER) - Primary Key: Unique identifier for each category.
- Genre (VARCHAR2(50)) - NOT NULL: Genre associated with the category.
- CategoryDescription (VARCHAR2(255)) - NOT NULL: Brief description of the category.

Relationship Between Tables:

- The Video table has a foreign key (CategoryID) that references the Category table, establishing a connection between videos and their associated categories.
- The Creator table is linked to the Video table through the foreign key (Username), indicating the creator of each video.

-Solutions: all your questions and the corresponding answers.
-Each answer should include
-A description of your answer/insights from your query results
-the SQL or PL/SQL code
-the query results

Question 1:

Which country (out of USA and India) has the higher average number of subscribers per creator?
Business Value: This question addresses the need to identify which countries produce the most followed YouTubers, providing valuable insights for content creators, marketers, and analytics teams to tailor strategies based on regional popularity.
Insights from Query Results: The results we got from this query showed that India has a higher number of people subscribing to different channels on Youtube. This is not surprising, as India has approximately 1.2 billion people while the United States has .4 billion people in population. What is interesting is that if you look at population to subscriber ratio, the United States actually has a higher percentage. So these results are true, but they also are misleading if you don't consider all the facts.
Results:
COUNTRY AVERAGESUBSCRIBERS

United States 23445833.333333333333333333333333333333
India 25658333.333333333333333333333333333333

```
-- Quesion 1

FUNCTION query_average_subscribers RETURN SYS_REFCURSOR IS
  rc SYS_REFCURSOR;
BEGIN
  OPEN rc FOR
    SELECT
      Country,
      AVG(NumOfSubscribers) as AverageSubscribers
    FROM
      Creator
    WHERE
      country = 'India' OR country= 'United States'
    GROUP BY
      Country
    ORDER BY
      AverageSubscribers;
  RETURN rc;
END query_average_subscribers;
```

Question 2:

Which creators have the most videos uploaded, and how many videos do they have?
Business Value: By determining creators with the highest video counts, this question offers insights into work ethic, production processes, and content scheduling, aiding marketers, collaborators, and analytics teams in understanding a creator's platform dynamics.
Insights from Query Results: This question aims to look at how many videos different creators have on their channel. Since we got only 1 result for the different channels, we wondered if there was an issue with the data sources we used.
Results:

USERNAME	NUMOFSUBSCRIBERS	TOTALVIEWS	COUNTRY	JOINDATE	TOTALVIDEOS
Conor Maynard	13500000	2750993392	United States	19-MAY-06	1000

```
-- Quesion 2
FUNCTION query_total_videos_per_creator RETURN SYS_REFCURSOR IS
  rc SYS_REFCURSOR;
BEGIN
  OPEN rc FOR
      SELECT
          *
      FROM
          Creator c
      ORDER BY
          TOTALVIDEOS DESC
      FETCH FIRST ROW WITH TIES;
  RETURN rc;
END query_total_videos_per_creator;
```

Question 3:

How many videos, on average, do creators upload in a week in each category?
Business Value: This question informs content creators and social media managers about their upload frequency compared to peers, facilitating better content scheduling and engagement strategies within specific video categories.
Insights from Query Results: This question gave us some interesting insight into how creators upload videos on Youtube. There are very few that have weekly uploads. The majority seem to upload far less frequently than we had anticipated.
Results:

CATEGORYID	GENRE	AVGVIDEOSPERWEEK
26	Howto & Style	0.0019404915912031047865459249676584734799 48
15	Pets & Animals	0.0016778523489932885906040268456375838926 17
22	People & Blogs	0.0015981403457794566322824349847450239721 05
10	Music	0.0015507011866235167206040992448759439050 7
1	Film & Animation	0.0014814814814814814814814814814814814814 81
27	Education	0.0013958682300390843104410943606923506420 99
25	News & Politics	0.0013568521031207598371777476255088195386 7
28	Science & Tech	0.0013427734375
20	Gaming	0.0012964563526361279170267934312878133102 85
23	Comedy	0.0012539184952978056426332288401253918495 3
24	Entertainment	0.0012302717142957573201167000597560546943 65
17	Sports	0.0012249897917517354022049816251531237239 69
29	Nonprofits & Activism	0.0010548523206751054852320675105485232067 51

```
-- Quesion 3
FUNCTION query_average_videos_per_week RETURN SYS_REFCURSOR IS
  rc SYS_REFCURSOR;
BEGIN
  OPEN rc FOR
      SELECT
          cat.CategoryID,
          cat.genre,
          COUNT(v.VideoID) / NULLIF(SUM(TRUNC((CURRENT_DATE - c.JoinDate) / 7)), 0) AS AvgVideosPerWeek
      FROM
          Video v
      JOIN
          Creator c ON v.Username = c.Username
      JOIN
          Categorys cat ON v.CategoryID = cat.CategoryID
      GROUP BY
          cat.CategoryID, cat.genre
      ORDER BY
          AvgVideosPerWeek DESC;
  RETURN rc;
END query_average_videos_per_week;
```

Question 4:

What is the highest ratio of comments to views, indicating strong audience engagement?
Business Value: Identifying what high engagement means and looking at the ratios guides content strategists, creators, and marketers in formulating effective strategies to enhance audience interaction and replicate successful engagement patterns.
Insights from Query Results: This query looked at the ratio of comments to views, and as we can see, there are significantly more views than comments. This shows that many more people are watching videos without leaving any sort of comment. People looking at these statistics should aim to work towards possibly getting more comments by attempting to engage the audience in different ways.
Results:

TOTALCOMMENTS	TOTALVIEWS	COMMENTTOVIEWRATIO
2707102	2954552075703	0.000000916247854374297247197739452678284292811 9

```
-- Quesion 4
FUNCTION query_comment_to_view_ratio RETURN SYS_REFCURSOR IS
  rc SYS_REFCURSOR;
BEGIN
  OPEN rc FOR
    SELECT
      SUM(v.NumOfComments) AS TotalComments,
      SUM(cr.TotalViews) AS TotalViews,
      SUM(v.NumOfComments) / SUM(cr.TotalViews) AS CommentToViewRatio
    FROM
      Categorys c
    LEFT JOIN
      Video v ON c.CategoryID = v.CategoryID
    LEFT JOIN
      Creator cr ON cr.username = v.username
    ORDER BY
      CommentToViewRatio DESC;
  RETURN rc;
END query_comment_to_view_ratio;
```

Question 5:

What is the average number of comments per video for creators in each country?
Business Value: Calculating average comments per video in each country informs content creators and marketers about audience engagement, offering insights into potential growth opportunities and the effectiveness of content in different regions.
Insights from Query Results: The query results show the difference in how countries interact with Youtube videos. Some countries have populations that leave more comments than others.
Results:

COUNTRY	AVGCOMMENTS
Pakistan	741.6
Philippines	841
India	2319.357142857142857142857142857142857143
nan	7291.928571428571428571428571428571428571
United States	8163.592105263157894736842105263157894737
Canada	10904.875
France	19914.5
Japan	20544
Brazil	35577
United Kingdom	43784.5

```
-- Quesion 5
  FUNCTION query_average_comments_by_country RETURN SYS_REFCURSOR IS
    rc SYS_REFCURSOR;
  BEGIN
    OPEN rc FOR
      SELECT c.Country, AVG(v.NumOfComments)  as AvgComments
      FROM creator c
      INNER JOIN video v
      ON c.Username = v.Username
      GROUP BY c.Country
      ORDER BY AVG(v.NumOfComments);
    RETURN rc;
  END query_average_comments_by_country;
```

Question 6:

Which creator has the highest engagement rate (likes and comments) per video?
Business Value: Determining creators with the highest engagement rates assists marketers, collaborators, and influencer marketing teams in identifying highly engaged creators for potential partnerships.
Insights from Query Results: From these query results we can determine which channels have higher engagement from viewers. A lot of the celebrity talk shows (such as TheEllenShow, The Tonight Show Starring Jimmy Fallon, The Late Late Show with James Corden) had rather high engagement compared to other smaller channels, which makes sense.
Results:
USERNAME AVGENGAGEMENTRATE
Ed Sheeran 1719197

```
-- Quesion 6
  FUNCTION query_engagement_rate RETURN SYS_REFCURSOR IS
    rc SYS_REFCURSOR;
  BEGIN
    OPEN rc FOR
        SELECT
            c.Username,
            (SUM(v.NumOfLikes) + SUM(v.NumOfComments)) / COUNT(v.VideoID) AS AvgEngagementRate
        FROM
            Creator c
        JOIN
            Video v ON c.Username = v.Username
        GROUP BY
            c.Username
        HAVING
            COUNT(v.VideoID) > 0
        ORDER BY
            AvgEngagementRate DESC
        FETCH FIRST ROW WITH TIES;
    RETURN rc;
```

Question 7:

Which two categories have the highest total views and the lowest total views globally?
Business Value: Identifying categories with high and low total views globally helps media practitioners align content with proven successful genres, optimizing content planning for broader reach and audience engagement.
Insights from Query Results: Animation had the lowest views while Entertainment had the highest number of views. This is in line with previous results (such as the previous question where talk shows had high engagement). It also makes sense as animation is frequently seen as something “for children”, so many adults ignore that category.
Results:

CATEGORYID	GENRE	TOTALCATEGORYVIEWS
1	Film & Animation	4712624489
24	Entertainment	1431009478566

```
-- Quesion 7
  FUNCTION query_top_and_bottom_categories RETURN SYS_REFCURSOR IS
    rc SYS_REFCURSOR;
  BEGIN
    OPEN rc FOR
      WITH RankedCategories AS (
        SELECT
            Categorys.CategoryID,
            Categorys.Genre,
            SUM(Creator.TotalViews) AS TotalCategoryViews,
            ROW_NUMBER() OVER (ORDER BY SUM(Creator.TotalViews) DESC) AS RankDesc,
            ROW_NUMBER() OVER (ORDER BY SUM(Creator.TotalViews)) AS RankAsc
        FROM
            Video
        INNER JOIN
            Creator ON Video.Username = Creator.Username
        INNER JOIN
            Categorys ON Video.CategoryID = Categorys.CategoryID
        GROUP BY
            Categorys.CategoryID,
            Categorys.genre
      )      SELECT CATEGORYid, genre, TotalCategoryviews FROM RankedCategories
      WHERE RankDesc = 1 OR RankAsc = 1;
    RETURN rc;
  END query_top_and_bottom_categories;
```

Question 8:

Which creators have the highest like-to-subscribers ratio for their videos?
Business Value: Analyzing the like-to-subscribers ratio aids creators and advertisers in gauging audience engagement, allowing for strategic decisions on content creation and targeted advertising.
Insights from Query Results: The results from this query were that there were anywhere from one subscriber for every five likes to one subscriber to every three likes depending on the

popularity of the channel. It was slightly surprising, because we had thought there would be more subscribers for a percentage of likes than it actually turned out to be.

Results:

USERNAME	LIKETOSUBSCRIBERRATIO
LuisFonsiVEVO	0.0589744705882352941176470588235294117647
EminemVEVO	0.0381050607287449392712550607287449392713
Ed Sheeran	0.0305444859813084112149532710280373831776
SelenaGomezVEVO	0.0288226388888888888888888888888888888889
SQUEEZIE	0.0282304972375690607734806629834254143646
Anitta	0.0237378488372093023255813953488372093023
Vogue	0.0211356390977443609022556390977443609023
The Late Late Show with James Corden	0.0185345070422535211267605633802816901408
SMTOWN	0.0150507523510971786833855799373040752351
Unbox Therapy	0.0126782383419689119170984455958549222798
Jake Paul	0.0112589705882352941176470588235294117647
Marques Brownlee	0.008626395348837209302325581395348837209302
BuzzFeedVideo	0.008345024875621890547263681592039800995025
FaZe Rug	0.007520506329113924050632911392405063291139
VanossGaming	0.007266046511627906976744186046511627906977
jeffreestar	0.006338301886792452830188679245283018867925
nigahiga	0.006297095238095238095238095238095238095238
Fueled By Ramen	0.006017297297297297297297297297297297297297
Lokdhun Punjabi	0.005870305343511450381679389312977099236641
Speed Records	0.005867529411764705882352941176470588235294
Linus Tech Tips	0.005761153846153846153846153846153846153846
CaseyNeistat	0.004566190476190476190476190476190476190476
T-Series Apna Punjab	0.004534624277456647398843930635838150289017
The Slow Mo Guys	0.004353469387755102040816326530612244897959
CollegeHumor	0.004027346938775510204081632653061224489796
The Q	0.003960597014925373134328358208955223880597
FailArmy	0.003813151515151515151515151515151515151515
Dude Perfect	0.00366937815126050420168067226890756302521
REACT	0.0033142
Reaction Time	0.003268344827586206896551724137931034482759
SonyMusicSouthVEVO	0.003202888888888888888888888888888888888889
First We Feast	0.0022648
Tasty	0.002221943127962085308056872037914691943128
Matt Stonie	0.002201595092024539877300613496932515337423
H2ODelirious	0.00218291044776119402985074626865671641791

The Tonight Show Starring Jimmy Fallon
0.002003108974358974358974358974358974
PowerfulJRE 0.0018725165562913907284768211920529801324 5
Jimmy Kimmel Live 0.001860261780104712041884816753926701570681
WWE 0.001846010416666666666666666666666667
Smosh 0.001742567049808429118773946360153256704981
Troom Troom 0.001697394957983193277310924369747899159664
TheEllenShow0.001669764397905759162303664921465968586387
The Infographics Show 0.001600833333333333333333333333333333
PewDiePie 0.001575864864864864864864864864864864865
Jass Records 0.001495748031496062992125984251968503937008
Markiplier 0.001453607954545454545454545454545455
Rosanna Pansino 0.001014166666666666666666666666666667
DALLMYD 0.000828308823529411764705882352941176470588
Atlantic Records 0.000815179856115107913669064748201438848921
TED-Ed 0.000811436170212765957446808510638297 87234
Brave Wilderness 0.000722248803827751196172248803827751196172
YRF 0.000705329670329670329670329670329670329670329 67
jacksepticeye 0.000674717607973421926910299003322259136213
5-Minute Crafts 0.000608938826466916354556803995006242197253
BRIGHT SIDE 0.000567011235955056179775280898876404494382
ABS-CBN Entertainment 0.000545769230769230769230769230769230769231
T-Series 0.000541787755102040816326530612244897959184
WatchMojo.com 0.000505418326693227091633466135458167330677
ABS-CBN News 0.000338940397350993377483443708609271523179
HUM TV 0.000207526501766784452296819787985865724382
CNN 0.000166447368421052631578947368421052631579
CrashCourse 0.000137094594594594594594594594594594595
HAR PAL GEO 0.000101771300448430493273542600896860986547
SET India 0.000025169811320754716981132075471698113207 55

```sql
-- Quesion 8
  FUNCTION query_like_to_subscriber_ratio RETURN SYS_REFCURSOR IS
    rc SYS_REFCURSOR;
  BEGIN
    OPEN rc FOR
      SELECT
        cr.Username,
        SUM(v.NumOfLikes) / NULLIF(cr.NumOfSubscribers, 0) AS LikeToSubscriberRatio
      FROM
        Creator cr
      JOIN
        Video v ON cr.Username = v.Username
      GROUP BY
        cr.Username,
        cr.NumOfSubscribers
      ORDER BY
        LikeToSubscriberRatio DESC;
    RETURN rc;
  END query_like_to_subscriber_ratio;
```

Question 9:

What are the trends in video engagement (likes, comments) during the first half of 2017 November versus the second half of 2017 November in the United States?

Business Value: Examining trends in video engagement during specific time periods helps creators plan content release schedules, enabling them to anticipate and compensate for potential fluctuations in engagement and income.

Insights from Query Results: This was interesting because the data flipped in the second half of the month. So in the first half, likes were lower than the second half, and comments were higher than the second half. We theorized that this could be in part because of the business of the second half of the month. When Thanksgiving hits, people panic because Christmas is right around the corner, and they can't be bothered to leave comments as frequently. Likes are easier when you are busier.

Results:

PERIOD	TOTALLIKES	TOTALCOMMENTS
First Half	2173336	350194
Second Half	2957331	238932

```
-- Quesion 9
  FUNCTION query_monthly_activity RETURN SYS_REFCURSOR IS
    rc SYS_REFCURSOR;
  BEGIN
    OPEN rc FOR
      SELECT
        'First Half' AS Period,
        SUM(NumOfLikes) AS TotalLikes,
        SUM(NumOfComments) AS TotalComments
      FROM
        Video v
        JOIN Creator c ON v.Username = c.Username
      WHERE
        v.PublishedOn BETWEEN TO_DATE('2017-11-01', 'YYYY-MM-DD') AND TO_DATE('2017-11-15', 'YYYY-MM-DD')
        AND c.Country = 'United States'
      UNION ALL
      SELECT
        'Second Half' AS Period,
        SUM(NumOfLikes) AS TotalLikes,
        SUM(NumOfComments) AS TotalComments
      FROM
        Video v
        JOIN Creator c ON v.Username = c.Username
      WHERE
        v.PublishedOn BETWEEN TO_DATE('2017-11-16', 'YYYY-MM-DD') AND TO_DATE('2017-11-30', 'YYYY-MM-DD')
        AND c.Country = 'United States';
    RETURN rc;
  END query_monthly_activity;
```

Question 10:

How does audience engagement in likes differ between YouTube creators in the United States and India?

Business Value: Comparing audience engagement between the United States and India provides valuable insights into regional preferences, helping creators and marketers tailor content and engagement strategies for specific geographic audiences.

Insights from Query Results: In this question we looked at average likes between two countries. The data we got from this was that the United States had a higher number of average likes than India.

Results:

COUNTRY AVGLIKES

United States 71821.763157894736842105263157894736842 1

India 35823

```sql
-- Quesion 10
  FUNCTION query_avg_likes_by_country RETURN SYS_REFCURSOR IS
    rc SYS_REFCURSOR;
  BEGIN
    OPEN rc FOR
      SELECT c.Country, AVG(v.NumOfLikes) AS AvgLikes
      FROM Creator c
      JOIN Video v
      ON c.Username = v.Username
      WHERE c.Country
      IN ('United States', 'India')
      GROUP BY c.Country;
    RETURN rc;
  END query_avg_likes_by_country;
```

Team: Describe your team members, and the contributions made by each member (who worked on which parts of the project)

In Milestone 1, our team unanimously opted for Group C, due to our lack of interest in wanting to do research on content creators and content consumers. Collaboratively, we brainstormed and defined columns and attributes for the three tables. Each team member actively participated in shaping the structure of our database, ensuring a comprehensive foundation for subsequent queries.

For Milestone 2, Carissa formulated questions 1, 9, and 10, while Ben contributed to questions 7 and 8. Sophia handled question 5, and Hannah did questions 2, 3, 4, and 6. Team members cross-verified and refined each other's responses.

In Milestone 3, we adopted a collaborative approach where each team member wrote the code for two assigned questions. Carissa worked on questions 1 and 2, Ben did questions 7 and 8, Sophia handled questions 5 and 6, and Hannah tackled questions 4, 3, and 10. We collectively tested and refined the code to ensure compatibility with our database, making adjustments as needed. For question 9, Ben and Hannah collaborated to develop the code collectively.

For developing the database, Ben took primary responsibility for setting up the GitHub repository, downloading the data from Kaggle, formatting the columns and setting up the packages. Carissa helped with testing the code for the queries completed in Milestone 2. Hannah and Sophia cleaned up the documentation/ for the questions as we had changed some of the questions to better fit our database and crossed checked the assignment rubric to make sure that all checkpoints were met.

Carissa took charge of the presentation creating an outline for what needed to be added and added information readily available. Ben added the results for the queries. Hannah and Sophia added any other information missing and reviewed the slide deck. We all individually worked on the slides that we were going to present to the class as well.

For the Final Report, Carissa added the screenshots of the queries from the Github page and wrote out all ten insights for the ten questions. Ben added the results to each question and

helped refine some of the queries for clarity. Hannah and Sophia went through and made sure the information was all correct and cleaned up the information.