# Project Background, Goals and Deliverables

Instagram has become a dominant platform for personal branding, marketing, and social interaction, with millions of users and businesses using it to connect with audiences, promote products, and build communities. Social media influencers on Instagram are individuals, businesses or even governmental agencies who have established a presence on the platform and use it to engage with their followers. They can range from individual content creators (e.g., Mr. Beast, Cristiano Ronaldo, Jennifer Lopez, or any of your favorite non-business influencers), to businesses trying to connect with their customers (e.g., Nike, Starbucks, Adidas), or organizations aiming to promote public initiatives (e.g., FDA, FTC, WHO).

One important success factor for social media influencers is the level of user engagement with their posts. Engagement with a post refers to any interaction that users have with a piece of content on social media platforms like Instagram. These interactions indicate how interested or invested users are in the content. On Instagram, engagement includes actions such as liking, commenting, and sharing the post. Social media influencers actively track their engagement and try different strategies to maximize their engagement.

Engagement is an important factor to all types of social media influencers for various reasons. First, likes and comments provide valuable feedback from the audience. By analyzing which types of posts generate the most interaction, influencers can gain insights into their audience’s preferences, interests, and behaviors, allowing them to create more targeted content. Second, Instagram’s algorithm favors content with higher engagement, meaning that posts with more likes and comments are more likely to appear in users’ feeds, on the Explore page, or in recommendations. This boosts the content’s reach, attracting new followers and potential followers. Finally, for businesses, while likes and comments don’t directly equate to sales, they are often the first step in the customer journey. High engagement on Instagram can lead to increased website traffic, higher conversion rates, and, ultimately, more sales. For these reasons, user engagement is an important metric when it comes to creating posts on social media. Social media influencers track user engagement in various ways to learn what content types and content posting strategies lead to higher engagement and use this information to design posts that can maximize their users’ engagement with their content.

Although influencers track their own user engagement on Instagram through the panels that Instagram provides to them to understand which posts lead to more engagement, the data that each influencer works with only comes from their own posts, which may not be adequate to make inferences about how to design highly engaging social media posts. One needs a more comprehensive analysis to be able to provide insights on what content types and strategies lead to more engagement.

Your job in this project is to analyze a big dataset of Instagram posts that includes two user engagement metrics (number of likes and comments a post received) to realize how social media influencers can design engaging posts. We can then share this information with social media influencers who are interested in understanding how to increase their number of likes and comments. At the end of the project you will deliver the following:

1. A written report: the audience of this report is the professor and the TAs. The goal of the report is to summarize the steps you took to analyze the data to respond to the specific questions posted by the professor. I will use this to understand whether you used the correct methodology and ran the correct analysis to come up with the 2nd deliverable.
2. A visual presentation of your findings: the audience for this deliverable is social media influencers who are interested in understanding how to increase the number of likes and comments on their Instagram posts.
3. The Python Code you used to generate the report and the visual presentation (if you used Python for creating it).
4. The PowerPoint file you are planning to use to present your project to the class on the last day of the class.
5. A peer review form where you rate your teammates on a scale from 1-10 on the extent to which they contributed to the project overall. I will take the average of these points and will use them to determine you final grades. To learn more about how I will incorporate peer-evaluations, check out the syllabus.

All items must be submitted on Canvas. If you have any questions about the project, you can contact the course TA or the professor.

You are only permitted to discuss this project with your team members, the professor and the course TA. Any discussion of project details or sharing code and solutions with students in the class or others outside of your team will be considered cheating. Additionally, you are allowed to use generative AI tools (e.g., ChatGPT) solely for generating Python scripts. Using these tools for any other purpose is prohibited and will be regarded as cheating. Violations will be subject to the University’s academic integrity policies **and will result in a zero for all team members for the entire project.**

# The Dataset

The available data set is provided by Corentin Dugué who scraped this data from Instagram. Read [this](https://towardsdatascience.com/predict-the-number-of-likes-on-instagram-a7ec5c020203) page to learn more about how the data was collected. The original dataset had about 30,000 records but since your computers, like mine, may not be able to process big datasets fast enough, I got a sample from the data set for this project and cleaned it up a bit. The dataset is available on Canvas (instagram\_data.xlsx) and has the following variables:

* LIKES: the number of likes the post received at the time of data collection
* COMMENTS: the number of comments the post received at the time of data collection
* TEXT: the text used as the post caption
* DATE: the date the post was created
* TYPE: what type of post it was (1 = PHOTO, 2 = VIDEO)
* USERS IN PHOTO: the number of users in the post
* LINK: the Link to the Instagram post

# Project Questions

## The main objective of the project is to identify the factors that contribute to generating more likes and comments on an Instagram post. However, you should focus on addressing the following questions. Responding to these questions helps you come up with better recommendations for social media influencers.

## Question 1

Run two linear regression analyses to predict the number of likes and comments based on the following variables:

1. Number of followers
2. Number of following
3. Month of posting
4. Day of posting (Monday, Tuesday, ...)
5. Timing of the post
   1. Morning (between 8AM and noon)
   2. Afternoon (noon – 4PM)
   3. Evening (4PM-8PM)
   4. Night (8PM to midnight)
6. Number of hashtags used in the post
7. Number of people tagged in the post
8. Type of post (video or photo)
9. Number of users in the post
10. Length of the post’s text

After running the linear regression analyses,

* Interpret how well each model performed
* Explain which variables from the above list had a statistically significant influence on the number of likes and the number of comments a post received
* For each variable that had a statistically significant effect on likes or comments, interpret the beta coefficient in terms of how much it increased or decreased the number of likes or comments.
* Determine which variable had the most important influence on the number of likes, and the number of comments.

Notes:

1. You must clean the data before running any analysis. See the report requirements for all the steps needed to clean the data.
2. As you notice, some of the variables listed above do not exist in the data, you must create those variables before running the analysis.
3. In this analysis, you predict two dependent variables: the number of likes and comments. In a regression analysis, you can only use one dependent variable. So, you must run a separate analysis for the number of likes and one for the number of columns. That’s why the question is asking you to “Run two linear regression analyses.”
4. When interpreting your results, explain which month, day of the week, or time in the day led to getting more likes and comments.
5. To create the variables related to the month, day and time of posting, Ask ChatGPT for the syntax to extract this information from the “DATE” column. Then dummy code them in the format I requested.
6. To determine which variable had the most important influence on each dependent variable, you must the “standardized beta coefficients (also known as beta weights)” for each independent variable. A standardized beta (also known as a standardized coefficient or beta weight) is a regression coefficient that has been standardized, meaning it has been scaled so that it is measured in terms of standard deviations rather than the original units of the variables. A standardized beta can help you compare the strength of the effect of each independent variable on the dependent variable. A higher absolute value of a standardized beta indicates a stronger impact on the dependent variable. You can ask ChatGPT for the code to perform this calculation. Once you have the standardized betas, sort the variables based on the absolute values of their coefficients. The variable with the highest absolute value will be the most influential variable in predicting the dependent variables.

## Question 2

The goal of this step is to help you develop your critical thinking skills and creativity. As you read above, the data set does not include many variables. However, as a data analyst, you can use your creativity to create variables from the available data that could potentially help you develop better models and gain more insights from the data. For example, I asked you to create a few variables from the data in Question 1 and use them in the regression analysis (e.g., month of posting, day of week, time of day, the length of the post, number of hashtags and number of people tagged in a post). In this question, your job is to use your creative skills to come up with three new variables that you could build from the available data and use in the regression analysis to see if you can improve the predictive power (Adjusted R-squared) of the model and if you can discover more variables that influence user engagement.

## Think carefully about the data and propose three variables that (1) you can logically create from the available data, and (2) based on common sense and your understanding of the business context, you believe could affect the number of likes and comments a post receives. Create these variables and include them in the regression analysis above to see if they improve the predictive ability of the model. In the regression analysis include the same independent variables used in Question 1.

## In your report, explain what your proposed variables are, why you think they may influence the number of likes and comments, and how you created them from the available data. After using them in the regression analysis, describe how well the model performed and whether you were able to improve its predictive power. Additionally, for the two new variables you created, explain whether they had a statistically significant influence on either of the dependent variables. If either of the effects was significant, interpret the coefficient in terms of how much a unit increase in the new variable increased or decreased the number of likes or comments.

Notes:

1. When running the regression use all input variables listed in Question 1 and add your new proposed variables. As you know, you must always check for multicollinearity before running linear regression, so after adding the two new variables, check for multicollinearity and drop variables that are highly correlated from the regression analysis.
2. You can dig deeper into the TEXT variable and use your creativity to create new variables from this variable. For example, I proposed using its length as a predictor. You can focus on the content (e.g., whether it started with a specific symbol or a word, or whether a specific symbol was used in the text or not).

## Question 3

Consider the variables used in Question 1. Repeat the same analysis with one difference: replace the “Day of posting” variable with a new variable named “Is weekend.” This variable will show whether the post was made during the weekend or during the weekdays. Run the linear regression and

* Compare the two models predicting the number of likes (one with day of posting which you ran in your response to Question 1 and one with “is weekend”). Which model performed better?
* Compare the two models predicting the number of comments (one with day of posting which you ran in your response to Question 1 and one with “is weekend”). Which model performed better?
* Do posts made on the weekend get more likes and comments compared to posts made during the week? If yes, how many more likes and comments do posts made on the weekend get compared to those made during weekdays?

## Question 4

While some social media influencers have a small number of followers, others have millions of followers. For the purpose of this question, let’s divide social media influencers to micro influencers (less than 50,000 followers) and macro influencers (more than 50,000 followers). Divide your data into two parts:

1. Data from micro influencers
2. Data from macro influencers

Your objective is to run separate linear regression models using data from micro and macro influencers, and then compare the results. This will help you determine whether the strategies for creating engaging posts should differ between micro and macro influencers.

Run two linear regressions to predict the number of likes and comments with the data from **micro** influencers. Next, run two linear regressions to predict the number of likes and comments with the data from **macro** influencers. Use the following variables in the prediction:

1. Number of following
2. Month of posting
3. Day of posting (Monday, Tuesday, ...)
4. Timing of the post
   1. Morning (between 8AM and noon)
   2. Afternoon (noon – 4PM)
   3. Evening (4PM-8PM)
   4. Night (8PM to midnight)
5. Number of hashtags used in the post
6. Number of people tagged in the post
7. Type of post (video or photo)
8. Number of users in the post
9. Length of the post’s text
10. If any oth the three new variables you proposed in Question 2 had a significant impact on likes and comments in Question 2, include them in this model.

Create a table with five columns: independent variable name, the coefficient for the effect of this variable on the number of likes using the micro influencer data, the p-value for the effect of this variable on the number of likes using the micro influencer data, the coefficient for the effect of this variable on the number of comments using the macro influencer data, the p-value for the effect of this variable on the number of comments using the macro influencer data.

Compare the p-values for the effect of each independent variable on the number of likes across the two models (micro and macro influencers data):

* Was there any independent variable that significantly influenced the number of likes for micro but not for macro-influencers?
* Was there any independent variable that significantly influenced the number of likes for macro but not for micro-influencers?

Compare the coefficients regarding the effect of each independent variable on the number of likes across the two models (micro and macro influencers data):

* Was there any independent variable for which the coefficients were different between the two regressions? Explain how this variable influenced the number of likes differently when the influencers were macro vs micro.

Compare the p-values for the effect of each independent variable on the number of **comments** across the two models (micro and macro influencers data):

* Was there any independent variable that significantly influenced the number of comments for micro but not for macro-influencers?
* Was there any independent variable that significantly influenced the number of comments for macro but not for micro-influencers?

Compare the coefficients regarding the effect of each independent variable on the number of comments across the two models (micro and macro influencers data):

* Was there any independent variable for which the coefficients were different between the two regressions? Explain how this variable influenced the number of comments differently when the influencers were macro vs micro.

# Project Deliverables

## Final Report (20 points)

The final report must be written in Times New Roman, 12 points with double-space. It must be less than 20 pages. All sources must be cited in the APA format. Do not include any Python code in the report. Also, wherever possible, use images and tables to summarize your findings. Do not copy and paste output from linear regression into the report. Add all coefficients and p-values to a table instead. The report must have the following sections:

* Introduction
* Methodology
  + Data
  + Data cleaning
  + Data transformation
* Analysis Results
  + Question 1
  + Question 2
  + Question 3
  + Question 4
* Conclusions and recommendations
* References

|  |  |
| --- | --- |
| Section | What you include |
| Executive Abstract | Format the abstract in this manner (300 word limit)   * One sentence that introduces the problem. * 3-4 sentences that explain your data, methodology and analysis. * 4-6 sentences explaining your most important findings. * 1 sentence to concluede. |
| Introduction | Briefly explain (300 word limit)   * What’s the problem. * Why is the problem important. * Who cares about this. * What is your plan for solving the problem. |
| Methodology | One paragraph that explains you followed the CRISP data analysis methodology. |
| Data | Introduce the dataset by explaining   * How was the data collected? * What variables exist in the data and what was their scale (categorical or continuous). List all variables that you used in all analyses you ran to respond to Questions 1-4. * How many records exist in the data. * Include the descriptive statistics for numerical variables in one table. * Add one table, list all categorical variables in the first column and their unique values in the second column. * Explain how many null values existed in each variable. * What are the independent and dependent variables in the analysis? |
| Data cleaning | * What were the challenges with cleaning this data? * What steps did you take to clean the data? Explain each step in detail. * How did you detect and deal with missing values? * How did you detect and deal with outliers? * Explain any additional steps you took to clean the data to run a regressin analysis? * You must take the following steps to clean the data:   + Read the data, inspect the variables, check out the top 5 rows to learn more about the data.   + Drop any variable that you do not intend to use in the model at all.   + After importing the data into Python, ensure that each column has the correct data type, as this impacts analysis and model training.   + Ask for the descriptives for all numerical variables that you intend to use in the model, check out the mean and the range of each variable and think about whether those statistics indicate an error in the data (e.g., the minimum number of followers is negative)   + Rename variables when needed. Sometimes, variable names in a dataset are unclear or overly abbreviated. Giving variables descriptive and meaningful names makes the data easier to interpret, especially when collaborating with others or preparing reports. Clear names also reduce the need to constantly refer to the data dictionary.   + For categorical data, ensure that categorical data is consistent (e.g., no variations in spelling like “Yes” vs “yes” or no extra values recorded).   + Recode variables if necessary. Sometimes, recoding variables into a new format (e.g., combining categories or grouping values) can improve model performance or interpretability. Sometimes, it is necessary for your model to recode values.   + Detect missing values and deal with them. Explain in detail how you detected missing values, how you decided to deal with them and why you made this choice.   + Detect outliers and deal with them. Explain in detail how you detected outliers, how you decided to deal with them and why you made this choice.   + Check for duplicate rows in the data. Drop the duplicates from the dataset.   + Standardize or normalize variables if necessary.   + For categorical variables, use dummy coding if necessary.   + Check out the correlation between all variables that you intend to use in the model. Create a correlation table and include it in your report.   + Check for multicollinearity before running a regression analysis and deal with variables that are highly correlated. |
| Data transformation | * What new variables did you create and why? * Did you dummy code any variable for analysis? Explain which variables and how did you create the dummy coded variables. * If you changed the existing variables in any way (e.g., changed their type, their scale, etc.), explain what you did and why. |
| Analysis Results | This is just a title. |
| Question 1 | * What analysis did you run for this question? * How well did the model perform? * What variables had a significant influcene on the dependent variables? * Interpret the coefficients (unstandardized or coeff in Python output) of each variable that had a statistically significant influence on the DVs in terms of how much they increased or decreased the number of likes and comments. * Which independent variable has the greatest impact on predicting the dependent variable? To determine this, you need to calculate the standardized beta coefficients (also known as beta weights). |
| Question 2 | * What new variables did you create? And why do you think these variables may influence the number of likes and comments? * What analysis did you run for this question? * How well did the model perform? * What variables had a significant influcene on the dependent variables? * Interpret the coefficients (unstandardized or coeff in Python output) of each variable that had a statistically significant influence on the DVs in terms of how much they increased or decreased the number of likes and comments. * Did the variables you added to the model improve the performance of the linear regression model compared to the model you ran in Question 1? |
| Question 3 | * What analysis did you run for this question? * How well did the model perform? * What variables had a significant influcene on the dependent variables? * Did replacing this variable improve the predictive performance of the model? |
| Question 4 | * What analysis did you run for this question? * How well did the models perform? * What variables had a significant influence on the dependent variables for micro and macro influencers? * Did the coefficients for the input variables differ between the two models? * Based on this analysis, can you make different recommendations about increasing engagement to micro and macro influencers? |
| Conclusions and recommendations | * Begin by writing one paragraph summarizing what you did in this project. * Explain what you found in data analyst terms (we found that variables X, Y, Z increase likes, comments, and X is the most important factor influencing engagement). Only include the highlights. * Explain what you found in the project in a language understood by the general audience (e.g., we found that posts made in the morning get more likes and comments than post made at night) * Based on your findings, propose recommendations for social media influencers who are interested in increasing their engagement (e.g., we recommend influencers to post during the day and avoid posting at night). |
| References | Include all references you used in the report in the APA format. |

## Python Code (10 points)

You must submit the Python code used to prepare the report and the visualization. The code must have enough comments to be understandable for the audience. So, use comments to explain the goal of each cell. Use heading to divide the code into understandable parts (e.g., data cleaning, Question 1, etc.)

Right before submission, close the Notebook you’re working with and reopen it. Click on “Run all” and make sure there are no errors. Codes that throw errors will lose 5 points.

**Visual Presentation (10 points)**

The goal is to create a digital and visual presentation of your findings that could be shared with social media influencers who are interested in increasing their engagement. You can choose between designing a PDF report with visual elements, an infographic, or a short video (2-3 minutes). You must let me know which one you decide on by Nov 21st. This will give you a great chance to build your resume as you will have an example of a real-world project you conducted that led to useful information that can be shared with social media influencers and has implications for many micro and mega influencers. You can share this on your LinkedIn profile, put a link on your CV, or post it on your personal website.

Since creating this deliverable requires much creativity, I will not provide much guidance to make sure I do not limit your creativity. However, note that:

# The audience of the visual presentation is social media influencers. So, keep that in mind when designing the visual elements.

# You are analyzing data collected from Instagram and are using likes and comments as engagement metrics. So, you cannot generalize your findings to other social media platforms (e.g., LinkedIn) or other engagement metrics (e.g., shares, clicks). So be very careful in your language when you are making a visual presentation. All information must be based on what you found in the analysis.

1. Although the consumers of the visual presentation are non-data-analysts, you must include a brief description of your methodology in the report. This will help people understand how you came up with these recommendations and what was your methodology. So, add one paragraph in lay language (a language that your grandparents understand) to explain how you came up with these recommendations.
2. As an example, you could provide information on:
   1. What’s the best time frame to post to get more engagement?
   2. What’s the best day to post to get more engagement?
   3. Should people post more on the weekends or during the week to get more engagement?
   4. Is the length of the post important in getting more likes and comments?
   5. Does the number of hashtag make any change in the number of likes and comments?

# Final Presentation (12 points)

You will give a brief presentation of your project on the last day of class. You will have 5 minutes to present your work, followed by 2 minutes for questions from the professor, TAs, or the audience. In work meetings, you often have a limited amount of time to present your work. This presentation will give you the opportunity to practice highlighting the most important aspects of your work to an audience in a concise timeframe. You can decide which parts to present but you must briefly cover the following:

* For the analysis you ran, describe how well your models performed and what you found when running the models to respond to the prompted questions. You can present this information in a few tables on your slides.
* What innovative variables did you come up with, and did they have a significant influence on user engagement?
* What recommendations do you have for social media influencers based on your findings?
* One slide (with three statements) where each student in the group adds one sentence explaining the most important thing they learned from this project.

Grading for the presentation:

* Grading will be based on:
  + Whether you covered the four points above.
  + How well you presented.
  + How well you designed your slides.
  + How well you responded to questions posed by the professor or the audience.
* If you do not attend the presentation, you will receive a 0 for the presentation grade regardless of how much you contributed to the project.
* I will set a timer for 5 minutes during your presentation and will stop you when the timer goes off. If you run out of time, each student will lose 4 points. So, practice before presenting to ensure you will finish on time.
* Reading from notes, slides or your phone is not allowed. You must practice your presentation to present from memory. Any student who reads off of the slides or notes will lose 50% of their presentation points (6 points).