

**Quiz: R Project Draft**

Course Code and Number: CIS 663

Course Section: 01-SP24

Name: Kaushal Guragain

Student ID: M00361014

**Instructions**

|  |
| --- |
| Please read the instructions below. Failure to comply with the instructions will result in a penalty ranging from 5% to 100% of the grade |

1. Use this document as a template for your response. In other words, type your student information and your answers into the document. Please do not delete anything from this document **[5% penalty]**
2. Please name the MS Word file that you send me as follows: “YourStudentId\_FirstName\_LastName\_CourseCodeNumber\_SectionNumber\_Project\_Draft\_SemesterYear.doc” For example, a student named John Smith with a student ID 00100001 from section 01 will have the following file name: 00100001\_John\_Smith\_CIS663\_Section01\_Project\_Draft\_Spring2024.doc **[5% penalty]**
3. Please submit the assignment to via GitHub before the specified deadline. **There will be a penalty for every day late (please refer to the syllabus). If you don’t submit your assignment properly – your assignment will be classified as “late” until it is properly submitted.**
4. No cheating please **[100% penalty]**
   * This is an individual assignment: you must complete this assignment on your own. Cases where two or more submissions have a larger overlap or when a student doesn’t seem to be knowledgeable in the assignment that he or she submitted may be investigated
   * If you use text from the Internet, textbook, etc. please make put an appropriate citation to make it explicit that you are using information from these sources (even in paraphrased form)

**Assignment**

This assignment involves creating a rough draft of a proposal for a research project in R. Your project needs to provide an answer to an important question or a solution to an important problem. There is no need to provide details. What matters is a rough understanding of what the project will be about and what kind of data, packages, and analysis techniques you will use for the project. It’s OK if you don’t have a specific idea for a particular aspect of the project. If this is the case, then just provide some ideas on what can be a possible way to go.

**Identify Problem and Motivate its Importance [4pts]**

This section should have two parts or paragraphs. The first part should clearly define the problem that you are trying to address. The smaller and the more specific the problem is, the better it is. The second part should explain why this problem is important. Feel free to use external literature to define the problem and motivate its importance. Please make sure all the sources are properly referenced using the APA format at the end of this report.

YouTube is one of the major players in the social media industry. It has millions of users who watch videos on the platform every day. According to the monthly estimate from backlinko.com, YouTube has over 2.49 billion monthly active users among with over 80 million are paid subscribers. This shows that YouTube has high monthly customers that business can target. They can use the media as a market place for selling their business ideas as it provides creator with free service to make videos and post in on the media for consumers to look at. However, it is not as easy to get the people to stream their videos and market their products or generate revenue from those videos. The higher the views the video has, the more money it generates for the creator. This is where the problem lies.

There is no concrete solution to estimating the category or type of people that will market the business ideas and product effectively on the platform.

If there was a statistical tool to help people know what type of videos, on which category, and of what length could get more views for them, it would be easier for any person, business, or a company to making customer-oriented videos to market their product efficiently. People could sponsor channels to market their customer oriented product if they strategized this properly.

**Literature Review [4pts]**

Read some articles/literature related to the problem area that you are working with and summarize what you have read in one paragraph. What have other people said about the problem? What are the things that are known or not known about the problem.

I read the paper Elango, Dinesh, Social Media Video Creators Monetization and Business on YouTube by Dinesh Elango. In the paper he talks about how advertisers will select the video for ads marketing and how the earning is distributed to the video creators. He also talks about possible conversation rates of views into revenue. Another interesting paper I came across was, Modelling and statistical analysis of YouTube’s educational videos: A Channel’s owners’ perspective. Some on the things that the paper talked about was, using moving average to analyze the trend of view per day for the channel, the correlation between video uploading activity, the age of channel and its popularity, classifying videos based on metrices like number of comments, subscribers, shares and likes. Zipf distribution and pareto law, devices used to watch videos, and more. All of the analysis that have been done with above matrices for content-based analysis and finding user behavioral patters in response to the videos they have watched or the playlist they have viewed. After reading the paper the things that are known about the problem are: YouTube’s massive user base, importance of views for revenue, challenges in generating revenue, lack of concrete solution for predicting the views on the video. Some not known things are, an accurate statistical tool for predicting views, Effective strategies for video making, strategic sponsorship of channels based on category and location.

**Research Question(s) [3pts]**

What specific question(s) are you trying to address with your project? Please list using bullet-point format. What are some of the R packages that you may want to use for this project?

* What kind of dashboard can you build based on the data and analysis?

The main question that my project will address is can a accurate statistical model be made for creating a tool that can predict views on a YouTube video based on different video metrices. If so, with how much accuracy can it predict the view count in response to the data from previous years.

Some packages that I am interested in include but not limited to the following:

1. ‘caret’: for evaluating predictive models for video views
2. ‘glmnet’: for handling multicollinearity, commonly found in YouTube video metrices
3. ‘randomForest’: to handle nonlinear relationships and interaction between video metrices and to predict views accurately
4. ‘xgboost’: for video view prediction (frequently used in predictive modeling)
5. ‘tidyverse’: to clean and prepare data for modeling and using visualizations
6. ‘caretEnsemble’: to combing predictions of multiple models and improve overall predictive performance
7. ‘forecast’: to analyze temporal patterns and build forecasting models based on historical data.

The dashboard will be a simple selection interface that when selected or input with values will generate a visual chart showing the predictive modeling results, Segmentation analysis across different metrices, performance comparison with historical data, etc.

**Theory [4pts]**

Without analyzing data, what do you see as a solution to your problem now? What do you think will happen once you analyze your data? What do you believe in now in relation to the problem area?

In my perspective, the best solutions might be to better manages the metrices like video length, category, location, video quality, frequency, and audience engagement. I haven’t analyzed the data yet but I think I will get a mathematical insight into the factors influencing views. People say a human can be wrong at times, but the math is the language of the universe. It is a proven fact. So, the generated statistical visual representation and data will help me to uncover video trends, get key business marketing insights, strategize business opportunities, etc. Right now, the views are pretty random across many categories. People are changing their trends fasters than ever before. So, it is difficult to say that one category of video based on certain metrices like video length may get more views than the other. An example to support this is, which many people spend their whole day scrolling through YouTube Short’s, it’s hardly as effective as let’s say Mr. Beast (a renowned content creator for YouTube) publishing a video which will get million of views in an hour.

**Data[4pts]**

What kind of data do you need for your research project? Where can you get it? How can you download and organize it?

The data I need for my research project will need to have high volume of data with many metrices involved as discussed before. The metrices might be ‘years’ to compare historical data, ‘channel title’. ‘channel category’, ‘view count’, ‘likes’, etc. I have found some relevant data for my project from Kaggle. One of which is <https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/> with column variables like ‘published year’ for historical data, and many more metrices.

The data I will need needs to be in an excel, CSV, or any database file where I can use R language with SQL or other languages to extract, transform and load data for my project.

**Methodology [4pts]**

How are you going to analyze your data? What kind of packages or statistical techniques are you planning to use?

My analysis will be a predictive statistical analysis based on the following packages and techniques:

1. ‘caret’: for evaluating predictive models for video views
2. ‘glmnet’: for handling multicollinearity, commonly found in YouTube video metrices
3. ‘randomForest’: to handle nonlinear relationships and interaction between video metrices and to predict views accurately
4. ‘xgboost’: for video view prediction (frequently used in predictive modeling)
5. ‘tidyverse’: to clean and prepare data for modeling and using visualizations
6. ‘caretEnsemble’: to combing predictions of multiple models and improve overall predictive performance
7. ‘forecast’: to analyze temporal patterns and build forecasting models based on historical data.

I might analyze my data on methodologies like, Exploratory Data Analysis (EDA) using scatter plots, linear regression model, perform cross-validation using Mean Squared Error, Statistical Tests based on the hypothesis.

**References [2pts]**

Please create an APA-style bibliography listing R packages and external sources you are using in the write-up. For example:

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

Elango, D. (2019, January 24). Social Media Video Creators Monetization and Business on YouTube. SSRN. Retrieved from <https://ssrn.com/abstract=3320253>

Samant, S., & Gautam, S. (2019). Modelling and statistical analysis of YouTube's educational videos: A channel owner's perspective. Computers & Education, 128, 145-158. <https://doi.org/10.1016/j.compedu.2018.09.003>