MSIS 615 Group Project

Dr. Shan Jiang

The purpose of group project is to reinforce your learning in MSIS615. You should be able to use only what we discuss in class to complete the project. You are not expected to use materials beyond those discussed in class for the group project unless you truly understand what you are doing. Please also refer to our discussions on academic integrity.

Should you use knowledge beyond what's discussed in MSIS615, e.g., use a package or library that we do not discuss in class:

- ***You must list them at the beginning of your presentation and include references to the source of the knowledge (e.g., the webpage that informed your coding or the course that taught you the knowledge).***
- ***If you use more than one technique that are not discussed in this class, I may ask you to present your project in person and explain your codes in an oral exam to make sure you truly understand your work.***

Introduction

For final project, you will develop an automated web crawler to collect data from a website of your choice, store them in a database, and perform data analytics using Python.

1) First, identify a website as your data source, then identify target data fields your team plans to collect. You should aim to collect as much data as possible, even if you do not initially expect to use some data fields for analysis. This is because retroactive collection could be time-consuming if you find that you are missing some needed data later on.

2) Set up a database to store your data.

a) You can use any database, including sqlite3, MS SQL Server, MySQL, MongoDB, etc. But note that Excel or a csv file is not a database.

b) Based on the data fields you identified from the website, design and create one or more tables that host your dataset to be collected.

c) After finalizing your database tables, develop the web crawler so that it directly inserts data into your database (instead of, for example, downloading files as a csv file first, then importing the csv file into the database).

3) Based on collected data in the database, perform some analyses to obtain insights. The types of analyses can include at least three of the following (but not limited to):

a) Descriptive analysis

b) Visualization

c) Regression

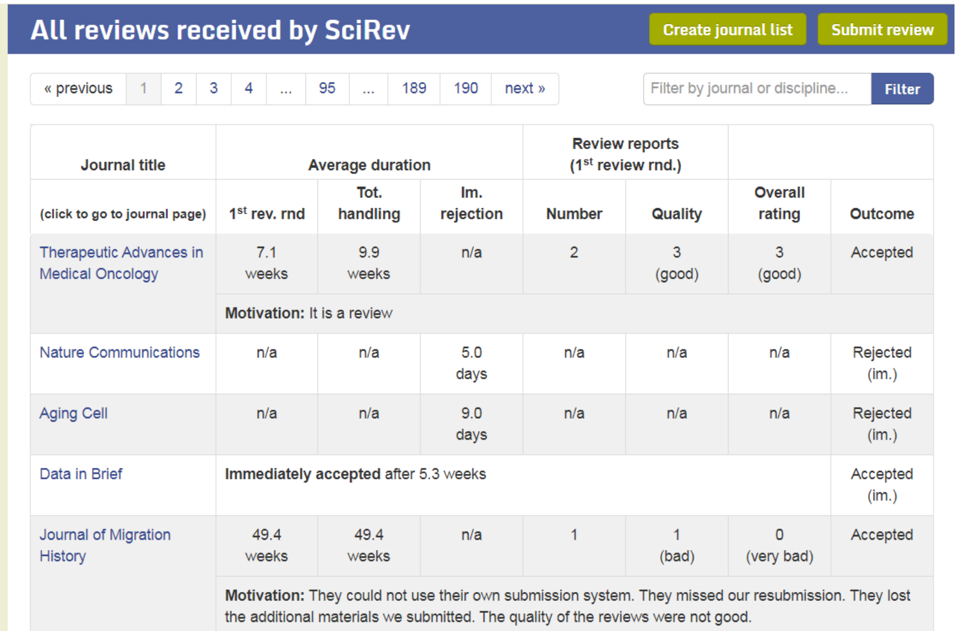
d) Sentiment analysis

e) Other text mining analysis

4) Present your work in a video presentation.

Example: SciRev.gov Analytics

Step 1: This example uses SciRev.org as the data source. The URL https://scirev.org/reviews provides a starting page for all reviews. All review information, including review text, journal title, average durations, review reports, overall rating, and outcome that are associated with the review process, are listed in 190 pages. Unfortunately, the website does not provide a button or APIs to download the data, which necessitates the development of a web crawler to automate the data collection.



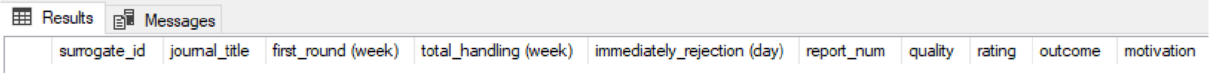
All reviews received by SciRev

Journal title	Average duration			Review reports (1st review rnd.)			
(click to go to journal page)	1st rev. rnd	Tot. handling	Im. rejection	Number	Quality	Overall rating	Outcome
Therapeutic Advances in Medical Oncology	7.1 weeks	9.9 weeks	n/a	2	3 (good)	3 (good)	Accepted
	Motivation: It is a review						
Nature Communications	n/a	n/a	5.0 days	n/a	n/a	n/a	Rejected (im.)
Aging Cell	n/a	n/a	9.0 days	n/a	n/a	n/a	Rejected (im.)
Data in Brief	**Immediately accepted** after 5.3 weeks						Accepted (im.)
Journal of Migration History	49.4 weeks	49.4 weeks	n/a	1	1 (bad)	0 (very bad)	Accepted
	Motivation: They could not use their own submission system. They missed our resubmission. They lost the additional materials we submitted. The quality of the reviews were not good.						

Step 2:

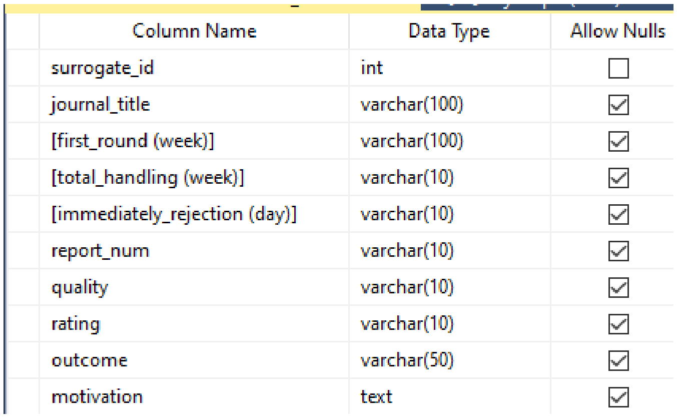
a. In this example, Microsoft SQL Server is used as the database to store the data.

b. To store the target dataset, including the identified data fields, the following empty table is created:



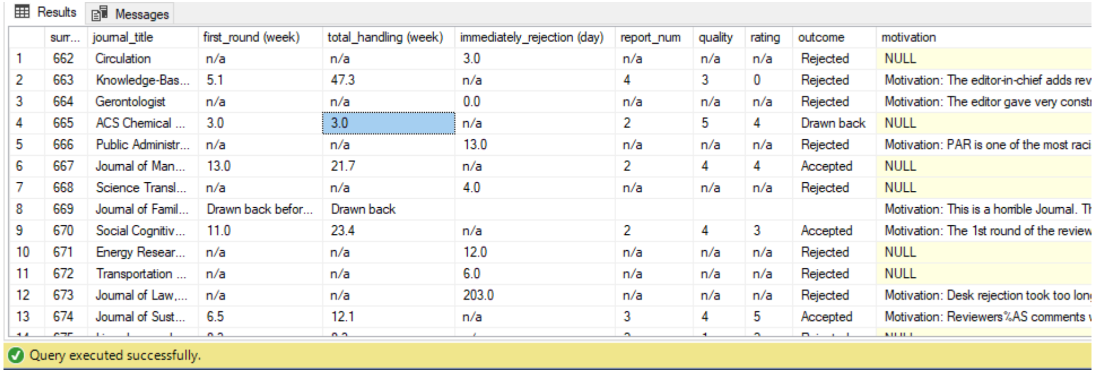
surrogate_id	journal_title	first_round (week)	total_handling (week)	immediately_rejection (day)	report_num	quality	rating	outcome	motivation

In order for each column to be able to store the corresponding data field correctly, the following data types are finally used after some trials and errors:



Column Name	Data Type	Allow Nulls
surrogate_id	int	☐
journal_title	varchar(100)	☑
[first_round (week)]	varchar(100)	☑
[total_handling (week)]	varchar(10)	☑
[immediately_rejection (day)]	varchar(10)	☑
report_num	varchar(10)	☑
quality	varchar(10)	☑
rating	varchar(10)	☑
outcome	varchar(50)	☑
motivation	text	☑

c. A web crawler is developed to collect all review data from the website. The web crawler directly inserts the collected data into the table above, resulting in 8,000+ rows (reflecting the reviews in the 190 pages) in the database.



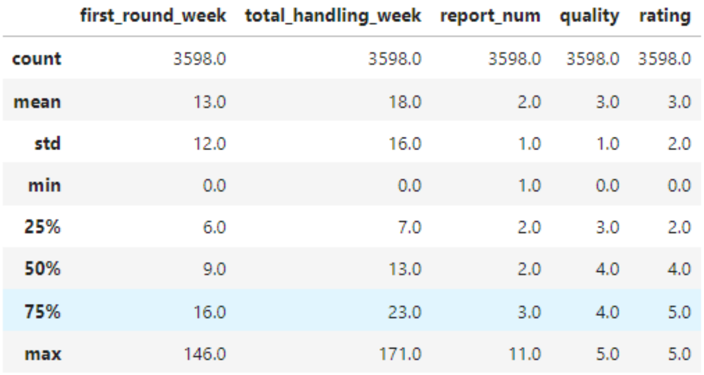
	surr...	journal_title	first_round (week)	total_handling (week)	immediately_rejection (day)	report_num	quality	rating	outcome	motivation
1	662	Circulation	n/a	n/a	3.0	n/a	n/a	n/a	Rejected	NULL
2	663	Knowledge-Bas...	5.1	47.3	n/a	4	3	0	Rejected	Motivation: The editor-in-chief adds rev
3	664	Gerontologist	n/a	n/a	0.0	n/a	n/a	n/a	Rejected	Motivation: The editor gave very constr
4	665	ACS Chemical ...	3.0	3.0	n/a	2	5	4	Drawn back	NULL
5	666	Public Administr...	n/a	n/a	13.0	n/a	n/a	n/a	Rejected	Motivation: PAR is one of the most raci
6	667	Journal of Man...	13.0	21.7	n/a	2	4	4	Accepted	NULL
7	668	Science Transl...	n/a	n/a	4.0	n/a	n/a	n/a	Rejected	NULL
8	669	Journal of Famil...	Drawn back befor...	Drawn back						Motivation: This is a horrible Journal. Th
9	670	Social Cognitiv...	11.0	23.4	n/a	2	4	3	Accepted	Motivation: The 1st round of the review
10	671	Energy Resear...	n/a	n/a	12.0	n/a	n/a	n/a	Rejected	NULL
11	672	Transportation ...	n/a	n/a	6.0	n/a	n/a	n/a	Rejected	NULL
12	673	Journal of Law,...	n/a	n/a	203.0	n/a	n/a	n/a	Rejected	Motivation: Desk rejection took too lon
13	674	Journal of Sust...	6.5	12.1	n/a	3	4	5	Accepted	Motivation: Reviewers%AS comments v

✓ Query executed successfully.

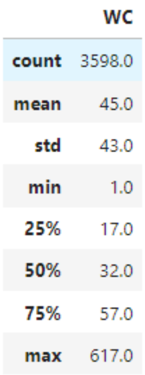
Step 3:

The entire table contains much missing information. The data is cleaned before analysis, resulting in 3,598 valid rows.

a. Descriptive analysis is performed, resulting in the following table. In addition, 1,498 reviews (41.6%) were for accepted papers 41.6%, and 2,100 reviews (58.4%) were for rejected papers. Average word count for the reviews was 45 words with a standard deviation of 43.0. The longest review had 617 words.

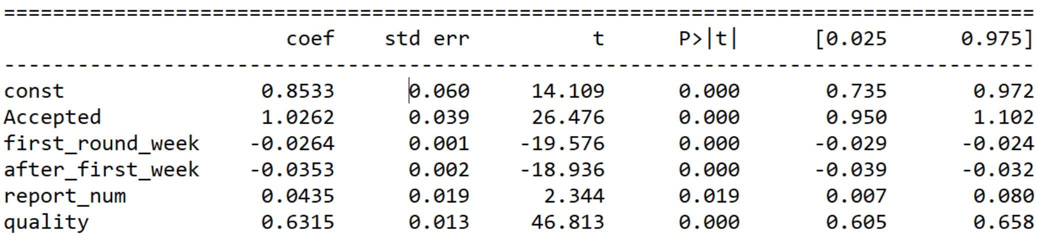


	first_round_week	total_handling_week	report_num	quality	rating
count	3598.0	3598.0	3598.0	3598.0	3598.0
mean	13.0	18.0	2.0	3.0	3.0
std	12.0	16.0	1.0	1.0	2.0
min	0.0	0.0	1.0	0.0	0.0
25%	6.0	7.0	2.0	3.0	2.0
50%	9.0	13.0	2.0	4.0	4.0
75%	16.0	23.0	3.0	4.0	5.0
max	146.0	171.0	11.0	5.0	5.0



	WC
count	3598.0
mean	45.0
std	43.0
min	1.0
25%	17.0
50%	32.0
75%	57.0
max	617.0

b. Regression analysis is performed using **statsmodels.api**, with overall review experience rating as the dependent variable (outcome variable) and other collected data fields as independent variables (input variables). Results are shown in the table below.

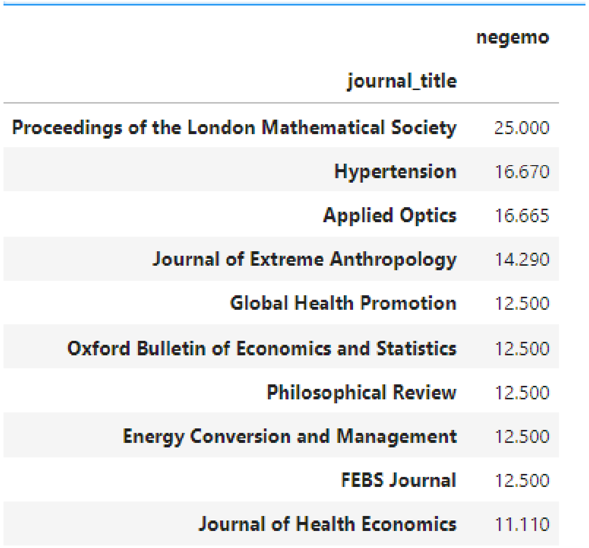


```
======================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const              0.8533      0.060     14.109      0.000       0.735       0.972
Accepted           1.0262      0.039     26.476      0.000       0.950       1.102
first_round_week  -0.0264      0.001    -19.576      0.000      -0.029      -0.024
after_first_week  -0.0353      0.002    -18.936      0.000      -0.039      -0.032
report_num         0.0435      0.019      2.344      0.019       0.007       0.080
quality            0.6315      0.013     46.813      0.000       0.605       0.658
```

The results show that paper acceptance, the number of reports, and review quality positively affect review experience, while the time spent on review processes (first_round_week, after_first_week) negatively affect the review experience. The p-values for all the variables are below 0.05 (5%), indicating a statistically significant correlation between the independent variables and the dependent variable.

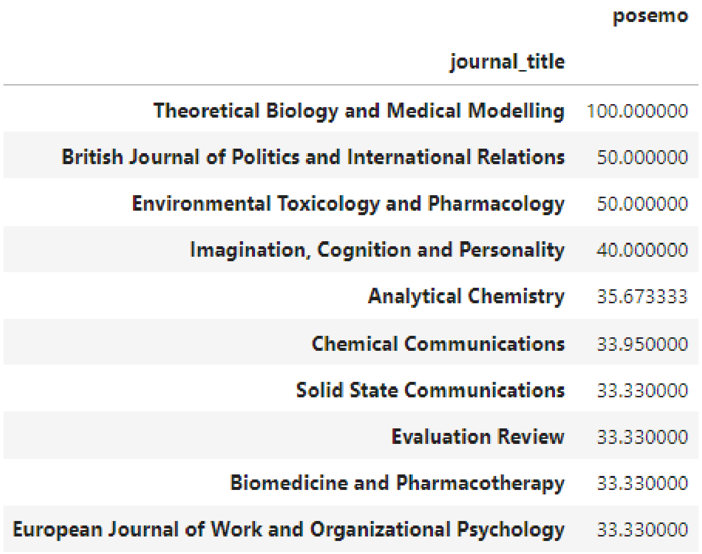
c. For the [review_text (motivation)] data field, sentiment analysis is performed. 2,658 reviews showed a positive attitude towards the review process, 1,045 reviews showed a negative attitude towards the review process, 810 reviews showed a mixed feeling of both.

The following journals were the top 10 journals with the most negative reviews:



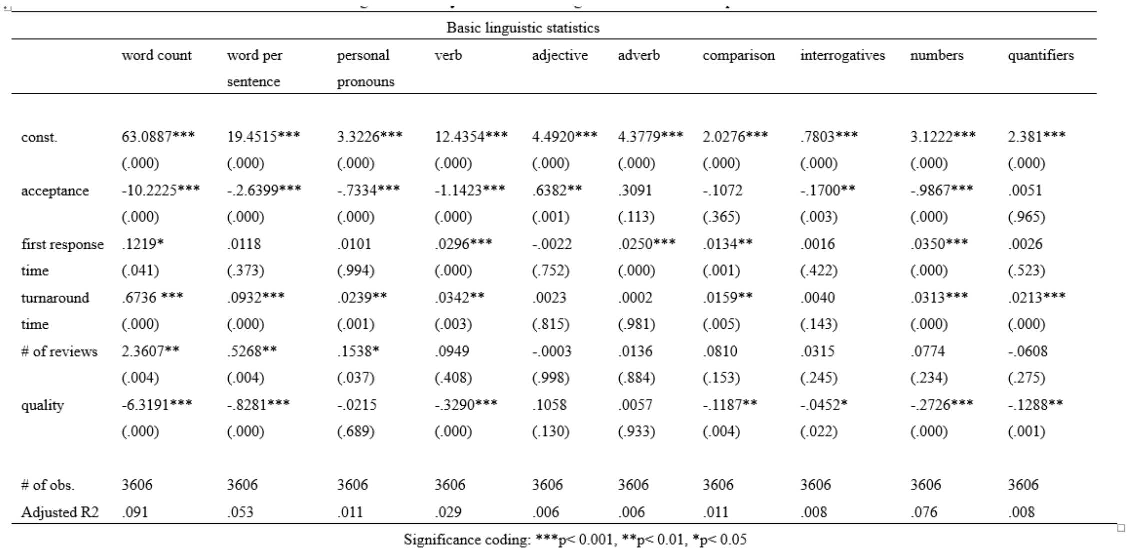
	negemo
journal_title	
Proceedings of the London Mathematical Society	25.000
Hypertension	16.670
Applied Optics	16.665
Journal of Extreme Anthropology	14.290
Global Health Promotion	12.500
Oxford Bulletin of Economics and Statistics	12.500
Philosophical Review	12.500
Energy Conversion and Management	12.500
FEBS Journal	12.500
Journal of Health Economics	11.110

In Comparison, the following journals were the top 10 journals with most positive reviews:



	posemo
journal_title	
Theoretical Biology and Medical Modelling	100.000000
British Journal of Politics and International Relations	50.000000
Environmental Toxicology and Pharmacology	50.000000
Imagination, Cognition and Personality	40.000000
Analytical Chemistry	35.673333
Chemical Communications	33.950000
Solid State Communications	33.330000
Evaluation Review	33.330000
Biomedicine and Pharmacotherapy	33.330000
European Journal of Work and Organizational Psychology	33.330000

d. Furthermore, additional text attributes (i.e., Part-of-speech tags) are extracted from the review texts to examine how review features will affect the use of these part-of-speech tags. The table below shows the result:



	Basic linguistic statistics									
	word count	word per sentence	personal pronouns	verb	adjective	adverb	comparison	interrogatives	numbers	quantifiers
const.	63.0887***	19.4515***	3.3226***	12.4354***	4.4920***	4.3779***	2.0276***	.7803***	3.1222***	2.381***
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
acceptance	-10.2225***	-.2.6399***	-.7334***	-1.1423***	.6382**	.3091	-.1072	-.1700**	-.9867***	.0051
	(.000)	(.000)	(.000)	(.000)	(.001)	(.113)	(.365)	(.003)	(.000)	(.965)
first response	.1219*	.0118	.0101	.0296***	-.0022	.0250***	.0134**	.0016	.0350***	.0026
time	(.041)	(.373)	(.994)	(.000)	(.752)	(.000)	(.001)	(.422)	(.000)	(.523)
turnaround	.6736 ***	.0932***	.0239**	.0342**	.0023	.0002	.0159**	.0040	.0313***	.0213***
time	(.000)	(.000)	(.001)	(.003)	(.815)	(.981)	(.005)	(.143)	(.000)	(.000)
# of reviews	2.3607**	.5268**	.1538*	.0949	-.0003	.0136	.0810	.0315	.0774	-.0608
	(.004)	(.004)	(.037)	(.408)	(.998)	(.884)	(.153)	(.245)	(.234)	(.275)
quality	-6.3191***	-.8281***	-.0215	-.3290***	.1058	.0057	-.1187**	-.0452*	-.2726***	-.1288**
	(.000)	(.000)	(.689)	(.000)	(.130)	(.933)	(.004)	(.022)	(.000)	(.001)
# of obs.	3606	3606	3606	3606	3606	3606	3606	3606	3606	3606
Adjusted R2	.091	.053	.011	.029	.006	.006	.011	.008	.076	.008

Significance coding: ***p< 0.001, **p< 0.01, *p< 0.05

** If you are interested, you can find the entire analytics in this paper: https://link.springer.com/article/10.1007/s11192-021-04032-8*

Grading

1. Project Functionalities (30 points)

Scores will be assigned to each requirement depending on the difficulty, quality of implementation, novelty, and amount of work.

1. Web Crawling (10 points): The students must demonstrate the ability to crawl a website.

An appropriate website is identified as the target data source. The data is spread in multiple web pages. Some websites may provide a download button or APIs (such as Twitter API) for data download. As long as you do not take advantage of such functionalities, you can still use these websites for the group project.

NOTE 1: If developing a web crawler is too difficult for you, you can still use Amazon.com as your target data source, but you need to make some changes (e.g., using a different product) and you'll lose up to 5 points for doing so.
NOTE 2: If you use any packages/libraries other than the Request and Selenium we discussed in class, make sure to credit the source of knowledge that inspire your coding.

2. Database (5 points) Set up a database to store your data.

1) (3 points) You should use any databases including sqlite3, MS SQL Server, MySQL, MongoDB etc. But note that Excel or a csv file is not a database. 6 points can be given if tables are correctly created and the data are successfully stored in and retrieved from the database.
2) (1 point) Your database should have at least five (more are encouraged) data fields (i.e., database columns, NOT including IDs) are collected from the target data source.
3) (1 point) At least 500 (more are encouraged) data records (i.e., database rows) are collected into the database.

3. Analytics (15 points): You are required to perform o*ne descriptive analysis* (including data visualization) and *at least two different types of in-depth analysis* covered in the course. Each analysis will earn you up to 6 points. This is where creativity and novelty will be rewarded.

> *NOTE: examples of in-depth analysis can be, but are not limited to regression analysis, other statistical analyses, machine learning, sentiment analysis, or other text mining analysis. Depending on the quality of the analysis, each analysis will be given 0-6 points. You are encouraged to do more different types of analysis, but the maximum number of points you can receive from the analysis part is 15. For example, in the SciRev.org example above, the following points may be assigned to each analysis:*

descriptive analysis	*- 3 points (descriptive analysis earn less points in general)*
regression analysis	*- 4 points (The analysis is simple but overall Ok)*
sentiment analysis	*- 5 points (good enough, but not outstanding)*
POS extraction + regression	*- 6 points (outstanding)*

> *The total from the Analysis module will be 15 points (because 3+4+5+5>15).*

The grading criteria for your analyses will be reasonable: Your own analyses don't need to be as good as the example to earn 6 points. However, technical correctness will not guarantee you full scores for analytics. To earn 6 points for your analysis, you must provide sufficient justifications for why an analysis is necessary, conduct the analysis correctly, and offer sensible interpretation of the results. Also note that you can perform same type of analysis no more than twice (e.g., running linear regression models on two sub datasets is OK, but the second one is unlikely to earn you 6 points unless it is perfectly done. A third similar linear regression won't early any points).

2. Presentation of your work (10 points)

Exactly how you record your *video presentation* is up to you. For example, you may use Zoom meeting recording or simply take a video of your group presentation using your cell phone. However, PowerPoint slides with inserted audio is **NOT** considered a video presentation.

The following shows the rubric for grading your presentations:

Rubric Item	**Explanation**	**Point**
Group work	Students should sign up for a group by the group formation deadline. The number of students must be no more than 4. Groups are expected to work through disagreements and conflicts. Points may be deducted for groups with membership changes after the group formation deadline.	2
Formality	Presentation slides are typo-free, grammatical error-free, and professionally formatted. All members dress appropriately for the video presentation (business casual)	3
Communication	Presentation length should be between 10-12 minutes. All group members are expected to be involved in the presentation. Students should present their topics in a way that everyone can understand, especially for audiences who are not familiar with the selected topic and/or analytic skills. Make sure to explain and justify your topic/questions, data source, and analyses you conducted.	5
Total		**10**

3. Submission of your work

Always check the course website for the project due date.

Files need to be uploaded as a group:

1) ***Source code***: Only upload .py files.
2) ***Databases***: If you use a local database (such as SQLite), upload the database file (such as .db database file). If you use a remote database (such as SQL Server), create a test account for the professor. Submit a file that includes all information on connecting to the database using the test account.
3) ***Presentation slides*****:** The slides should summarize why your project is interesting and worth doing, what data source you crawled, what database, tables, and fields you use to store data, what analysis has been done, and what are your main findings. *Please make sure to highlight the technical achievements that you are most proud of.*
   If you use a tool other than Microsoft PowerPoint to build the slides, please make sure the file you submit can be opened with Microsoft PowerPoint.
4) ***Video presentation***: Please either upload a file or a link to your video stored online. If your video file is too large to be submitted through course website, you **MUST** take advantage of the Microsoft OneDrive that you can use for free as a UMB student and share the link to the file. See this link for mor information on OneDrive through UMB:
   https://www.umb.edu/it/admin_systems/onedrive
   Please do NOT use any other file sharing services such as Google Drive and Dropbox. Nor should you upload your video to video sharing sites such as Youtube and share the link.

File to be submitted individually:

5) ***Peer evaluation form***: Students must upload a peer evaluation form to receive a grade for group project.

4. Individual grades adjustment

Each group will receive a group grade, but grades for individual students will be adjusted by the professor based on peer evaluation forms individually and confidently submitted by group members.

S***tudents who fail to submit their peer evaluation forms in time may lose partial or all points for their individual project grade.***