**MISM3403 Project Report**

**PR Package Management Database and Insights**

Keetae Kim, Peiying Li

**Topic and Business Questions**

We have built a database for a PR package management platform that PR managers can use to efficiently send products to targeted content creators and generate analytical insights. We answer the following three business questions of interest: (1) Which Tiktokers are paid the most? (2) Which Tikoker has the highest quantity of sales? (3) What is the best-selling company?

[Adobe](https://www.adobe.com/express/learn/blog/content-creator) defines content creators as “someone who creates entertaining or educational material to be expressed through any medium or channel.” For example, creative individuals could become “Youtuber, ” “Tiktoker,” or an influencer on Instagram. Through these social media platforms, they can create personalized content and gain income through view counts, advertising, and brand contrasts. According to [Cision PR Newswire](https://www.prnewswire.com/news-releases/global-digital-content-creation-market-size-expected-to-reach-38-billion-by-2030-301381125.html), “the Global Digital Content Creation Market size was estimated to be US$ 11 billion in 2019 and is expected to reach US$ 38.2 billion by 2030 at a CAGR of 12% during 2020 and 2030.” Sending products to be featured by content creators on their platform can potentially increase the sale of the product, while the process itself could be dreary including finding popular content creators, getting their addresses, tracking the shipment status, etc. A blog article from [Statusphere](https://brands.joinstatus.com/how-to-send-pr-packages-to-influencers) suggests that “ the process as a whole can take on average 70 hours of work.”

Given the business potential and to help PR managers market their product more efficiently, we have built a tool that would assist them with sending packages and provide analytical insights. The potential users of our platform are a PR company or any marketing team of a company. The database of such a tool, stores information such as the content creators’ address, contact information, relative product information, whether or not a product has been sent, on what date the product is sent, etc. This information supports the “basic” functionalities of the database that enable the PR managers to automate the package sending process by easily looking up the address of a content creator. While the extended functionalities include tracking how many sales have been made since the product is featured. The supported functionalities and the business questions mentioned above will allow PR companies to make more targeted marketing decisions.

**Data**

The first dataset that we use is the [*TikTok Trending Videos*](https://www.kaggle.com/erikvdven/tiktok-trending-december-2020?select=trending.json)from Kaggle which stores the first 1000 trending videos scraped from TikTok. The dataset is in JSON format and includes variables such as content creators ID, video description, play count, share count, hashtags, etc. This dataset is used to generate the content creator (Tiktoker) information and for analyzing the creator’s popularity. This data does not contain information such as “Tiktokers’” addresses and contact information. Therefore, we generate “placeholder” data for the purpose of this project. We also use a [helpful notebook](https://www.kaggle.com/erikvdven/tiktok-some-python-magic-in-a-notebook) that will help us with processing the data.

The second dataset that we use is the [*2021 Fortune 1000*](https://www.kaggle.com/datasets/winston56/fortune-500-data-2021)from Kaggle which stores information on the top 1000 companies from the Fortune magazine in 2021. The dataset is in CSV format and includes variables such as the company name, the revenue a company generates, the sector a company belongs to, etc. This dataset was used to create the company information affiliated with the product, the PR managers, and the products as it was impossible for us to scrap real information on the Tiktokers. This dataset, along with additional sale information generated (explained in the data creation below) will answer business question #3 on the best-selling companies.

The third dataset that we use is the [*2020 Amazon Product*](https://www.kaggle.com/datasets/promptcloud/amazon-product-dataset-2020)from Kaggle which stores information on 30K Amazon products. The dataset is also in CSV format and includes variables such as the product name, product details, category, etc. This dataset was used to create the products that may be sent to the content creators. Similar to the company data, it was impossible for us to find real product sending information online. Therefore, we used this dataset to generate our own product catalog. This dataset, along with the Tiktoker information and additional sale information generated (explained in the data creation below) will answer business questions #1 and #2 on the best-selling Tiktokers.

**Information Quality**

Because of the functionalities that we aim to support with our database, it was challenging to find a dataset that fulfills all the needed information. Therefore, we had to pull information from the different sources mentioned above.

For the *TikTok Trending Videos* dataset, the author mentions how the data was generated using his personal login information which might result in personalized trending videos. From this data, we were able to extract the top Tiktokers and calculate the popularity based on the view counts, download counts, and like counts. However, the data does not include the Tiktokers’ addresses, which are important for package sending. There is no data cleaning needed other than to load the JSON format data into a data frame in a Python script, and to select the columns that we are interested in. Please see the code script for details.

Similarly, the *2021 Fortune 1000* dataset includes the necessary information that we are interested in, such as the sectors that the companies belong to. However, it does not include detailed addresses and contact information of the companies. There is no additional data preprocessing needed other than to load the CSV format data into a data frame in a Python script and to select the columns that we are interested in using. Please see the code script for details.

Lastly, the *2020 Amazon Product* also includes information that we are interested in such as the category and the price of the product. In contrast with the other two datasets, this dataset needs additional cleaning as some product prices were malformed. For example, most of them included the “$” which needs to be removed so that it could be recognized as an integer or a float.

We mainly used these three datasets to generate the “tiktoker” table, the “product” table, and the “company” table. For the tiktoker and the company table, we use various Python libraries, such as “[Faker](https://faker.readthedocs.io/en/master/)” to generate additional information such as addresses. We also wrote customized functions to generate contact information such as phone numbers. To support the functionalities of our database, we also generated data for tables such as “feature” and “sale,” which will be explained in later sections. Please refer to the code script for details on data generation and preprocessing.

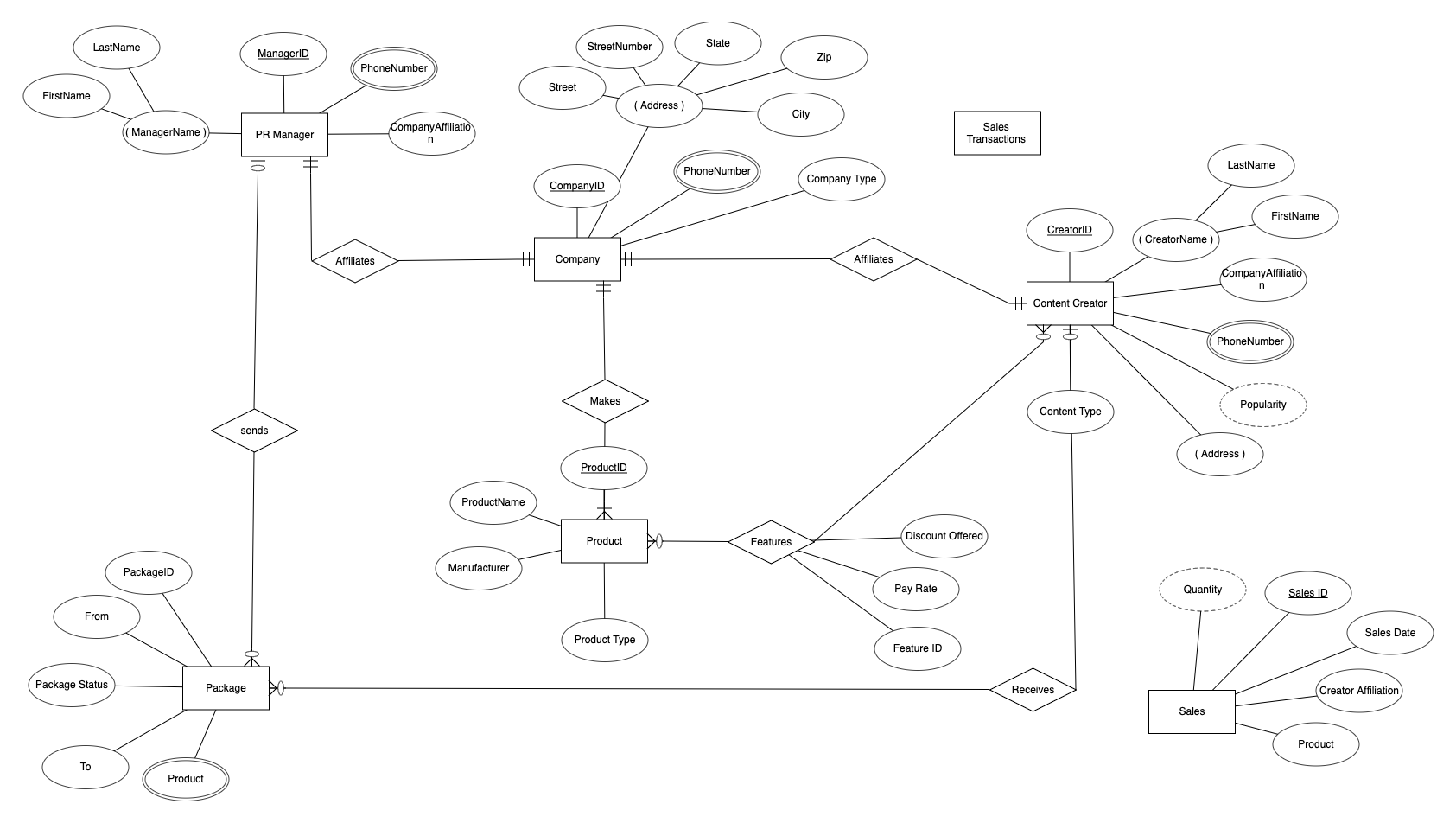
**Data Management Process and Model**

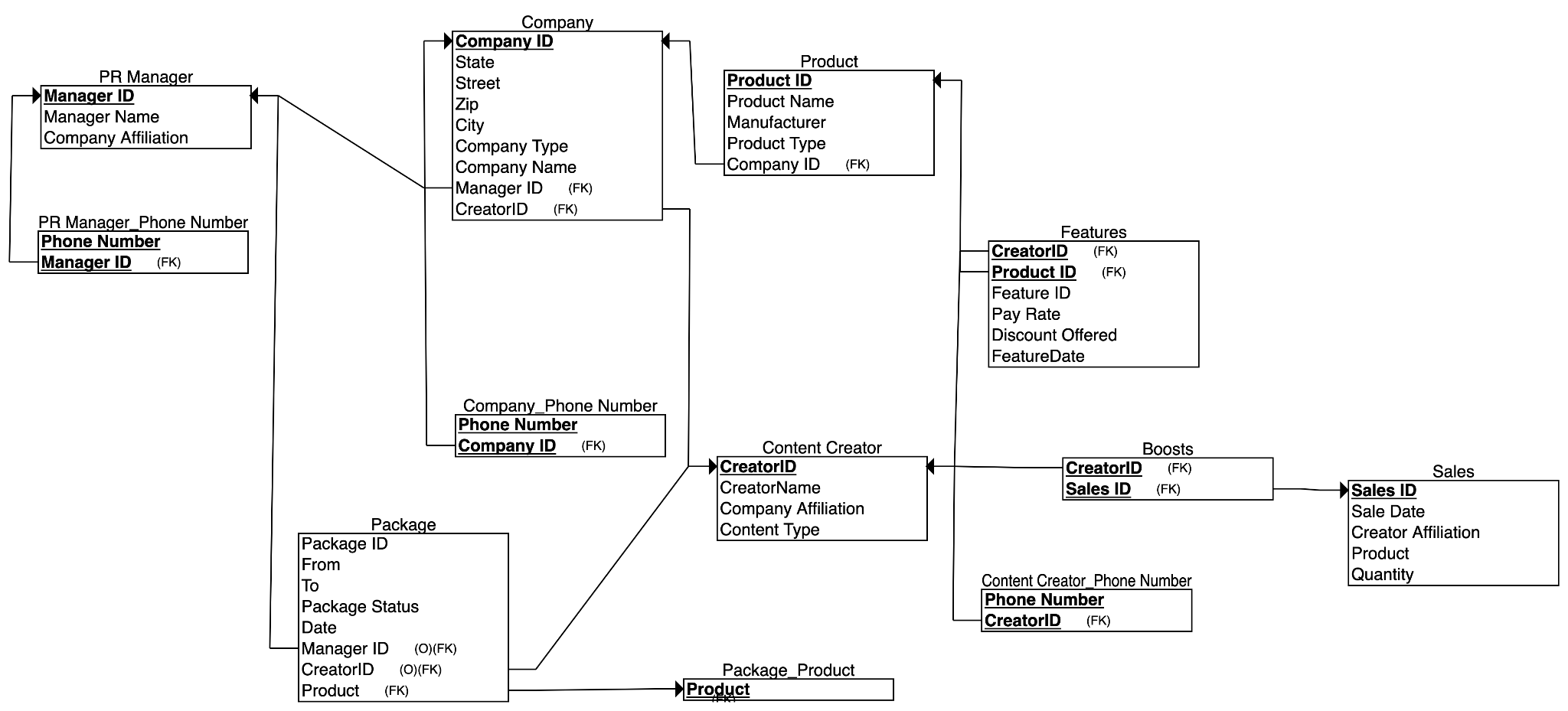
Users of the PR package management platform can send packages to content creators and track the status of the packages, the features on the products, the sale of the products and the payment information to the content creators . The content creators in our data models are the TikTok influencers. These “Tiktokers” and the users (the PR managers themselves) are affiliated with different companies. Ideally the content creator will boost the sales of the products included in the PR packages by featuring the products on their social media channel (Tiktok). Each PR manager will have a unique manager ID along with their name, company affiliation ID, and phone number. For content creators, each of their names, address, phone number, company affiliation, and type of content will be stored in addition to their popularity which will be calculated. This will allow PR managers to send packages to the right influencer in order to boost the sales of the affiliated product. The database also stores data on packages, specifically on what the packaged product is, status of the package, the manager's ID who is in charge of sending the product, as well as the receiving content creators ID. The product table on the other hand will store information such as the product name, product company, type of product, and a unique ID for each product. The database also stores the sale transaction of each featured product, including a unique sales ID, date, and the creator affiliation information allowing users to interpret how much a content creator has boosted sales. Last but not least, the feature’s info on which product is featured by which Tiktoker will also be stored with a unique feature ID, featurebycontentcreatorID, product ID, discount rate, and the pay rate to the content creator for each feature. This is essential to see how much the Tiktoker will make from featuring a product. For each feature, a Tiktoker will receive a flat rate and they can feature the product multiple times. For each affiliated sale, the Tiktoker will receive 0.05% of the product price.

The PR Manager can send zero or more packages showing an optional 1:M cardinality.

Content creators can potentially boost sales with a m:n cardinality. Both these entities are optional as a content creator may not increase sales or can decide to not showcase the product at all. The company is affiliated with both the PR manager as well as the content creator with a one (and only one) cardinality. Content creators are responsible for boosting sales with an optional m:n cardinality. Packages are received by the content creators with a M:1 cardinality.

Please refer to the database for the data types of each attribute shown in the ERD and the relational schema below.





**Results**

*Data Table Creation*

To support the functionalities of our database, we have created the following tables and imported our data into our database using Oracle developer server:

| Table Names | PR Manager | Package | Package  Status | Tiktoker | Company | Product | Feature | Sale |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rows | 200 | 1000 | 5 | 802 | 1000 | 1000 | 149 | 827 |
| Columns | 4 | 6 | 2 | 10 | 8 | 5 | 6 | 6 |

As mentioned earlier, we created the Tiktoker, Company and Product table with a combination of real data and randomly generated data in Python. For the rest of the tables, we generated the data with several Python libraries, including “[Random](https://docs.python.org/3/library/random.html)” and “[Faker](https://faker.readthedocs.io/en/master/).” We also made sure that the generated tables followed the constraints and relationship mappings from our data model. After we generated the data, we imported the tables above into the Oracle developer server to conduct data analysis and to answer the business questions. Please see the code script for detailed steps on data generation.

*Date Retrieval and Answers to Business Questions*

1. Which tiktokers are paid the most?

The total pay rate for each Tiktoker to feature a product is calculated as the sum of the flat feature rate of the product and the total pay rate they receive for each product sale associated with them. We calculated the pay for both the flat feature rate payment and the sale payment separately, stored those as two calculation tables, then aggregated the two numbers for each Tiktoker to calculate their final pay. The SQL code is as follows:

*select creatorid, sum("Total Features" + "Total pay") as "Total"*

*from featurestemp*

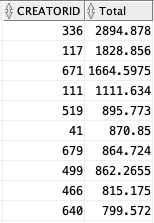
*inner join totalpay*

*using (creatorid)*

*group by creatorid*

*order by "Total" DESC;*

The top 10 top-paid Tiktokers are the following:



1. Which Tikoker has the highest quantity of sales?

To find out the TikToker with the highest amount of sales, we used the Sales table which includes the individual transactions and the purchased quantity. We grouped by the creatorID and sum the total quantity of sales. We use the following SQL code to answer this question:

*select creatorname, sum(quantity) sales*

*from sales*

*inner join tiktokers*

*using(creatorid)*

*group by creatorname*

*order by sales DESC;*

The top 10 Tiktokers that have the most quantity of sales are the following:



1. What is the best-selling company?

To answer this question, we used the Sales, Company and the Product table. We found out the manufacturer information that is associated with a product by joining the three tables. Then we summed the selling quantity while grouping by the manufacturer(company) name. We used the following SQL code to answer the question:

*select company.companyname, company.companyid, sum(quantity) "Number of Sales"*

*from product*

*join sales*

*on product.productid = sales.productid*

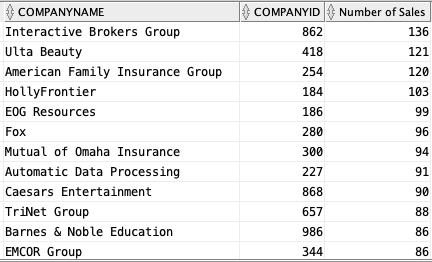
*join company*

*on product.manufacturer = company.companyid*

*group by companyname, companyid*

*order by "Number of Sales" DESC;*

The top 10 selling companies are as follows:



To better understand the results from these business questions and to generate additional insights, we created visualizations included in the Appendix. Users of our database could also conduct simple information lookups if they are interested in the top-selling content creator and/or company. Please keep in mind that the sales information from the data is not real (for demonstration of the functionalities of our database only).

**Additional Data and Analysis**

If we had more time, we would collect more viewer and buyer information. Understanding the demographics of the viewers watching the content creators could also allow us to conduct more targeted marketing based on our viewer and/or customer analysis. We would also conduct research on predicting future viral content trends based on the content information (that may not be used in this project). Predicting potentially viral content and content creators could open up numerous opportunities such as sending PR packages to new and upcoming creators that are more likely to make sponsored videos with lower costs. Another addition to this research would be extending to more platforms such as Twitter or Snapchat and expanding the database to include more countries to access a larger international pool because product preferences may vastly vary around the world.

**Contributions**

*Keetae*: Reference finding, database implementation, research and summarization of external materials

*Peiying:* Select proper dataset, data retrieving and processing, data analysis and visualizations

*Joint Contribution:* Database model design, draft slides and present, draft report

**References**

*Topic Reference*

[1] [The Importance of PR Packages](https://www.bu.edu/prlab/2022/03/07/the-importance-of-pr-packages/)

[2] [Global Digital Content Creation Market Size Expected To Reach $38 Billion By 2030](https://www.prnewswire.com/news-releases/global-digital-content-creation-market-size-expected-to-reach-38-billion-by-2030-301381125.html)

[3] [Influencer Marketing Platforms vs. DIY Influencer Marketing](https://brands.joinstatus.com/influencer-marketing-platforms)

[4] [How to Send PR Packages to Influencers at Scale](https://brands.joinstatus.com/how-to-send-pr-packages-to-influencers)

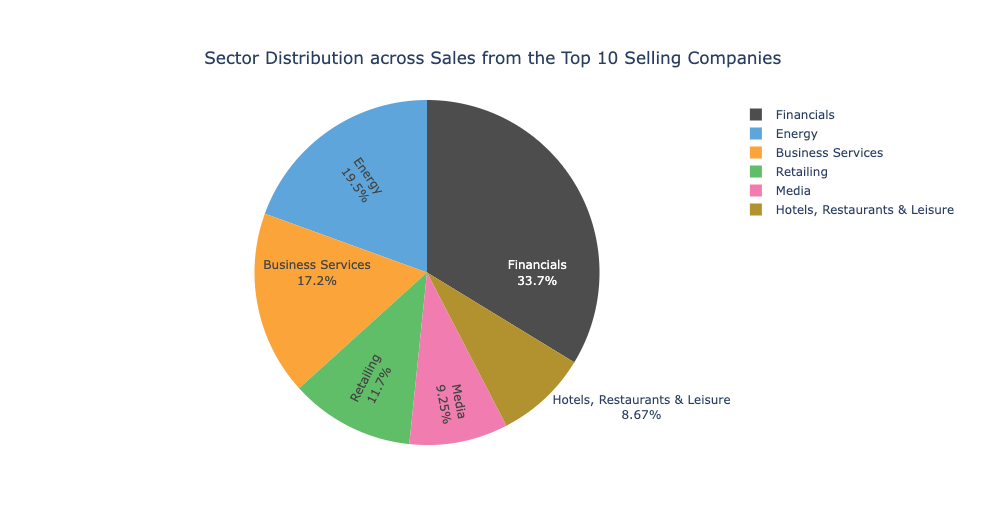
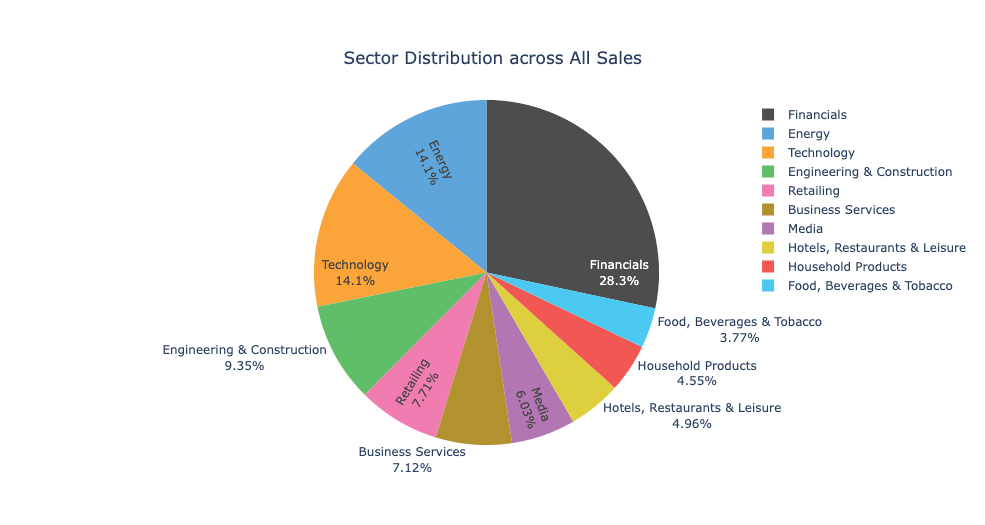
[5] [The Importance of PR Gifts for E-commerce Brands](https://www.found.co.uk/blog/importance-gifts-ecommerce-brands/#.YmdBF9OZNhE)

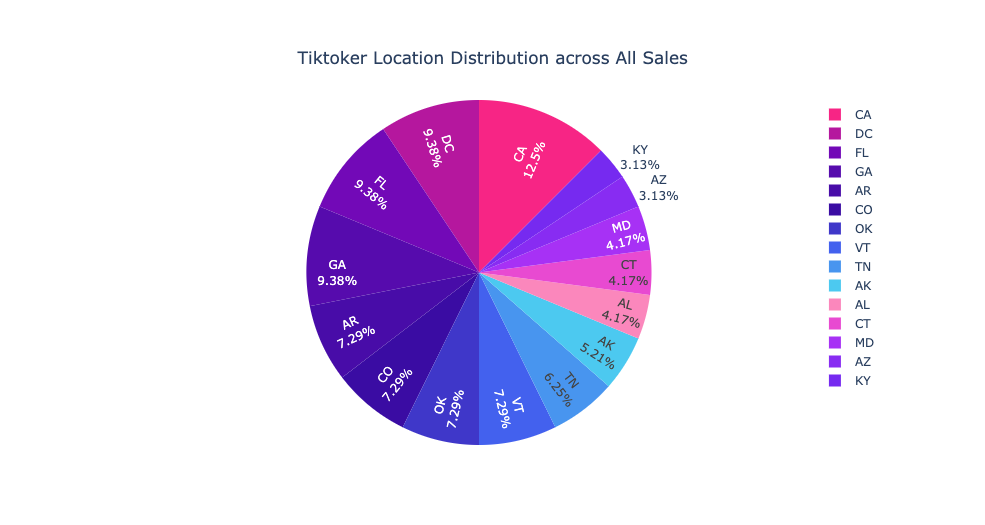
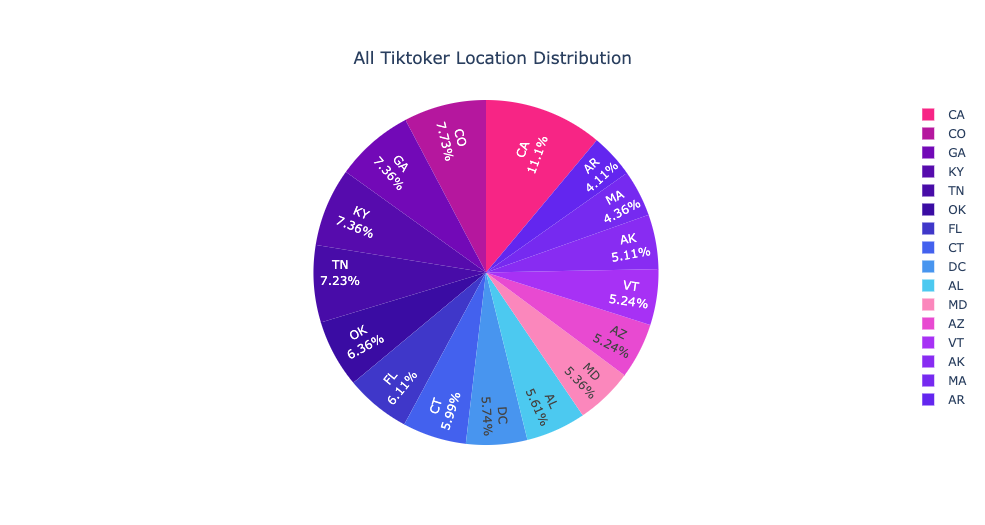
*Coding Reference*

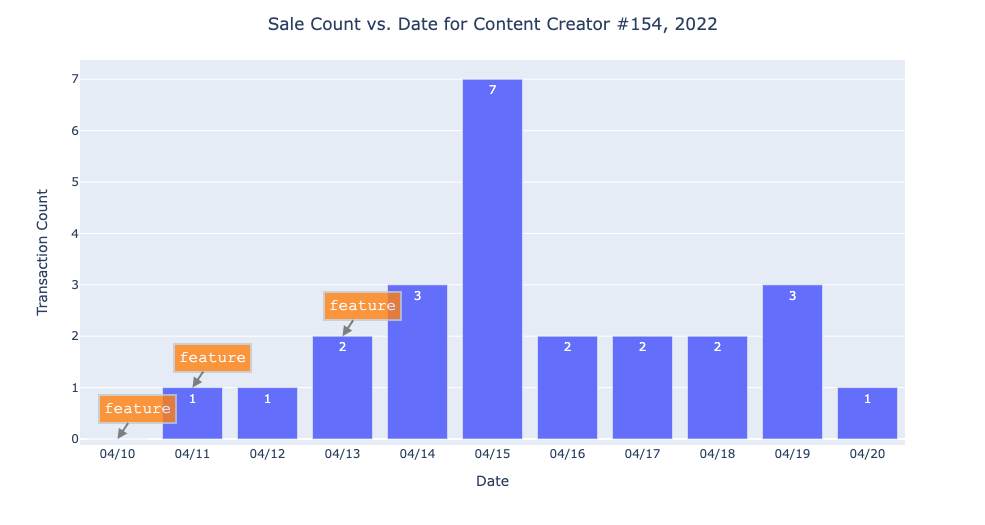
[1][TikTok - Some Python magic in a Notebook](https://www.kaggle.com/erikvdven/tiktok-some-python-magic-in-a-notebook)

**Appendix**

Please find below the additional visualizations that we generated based on the results from the three main business questions.





****