**САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ**

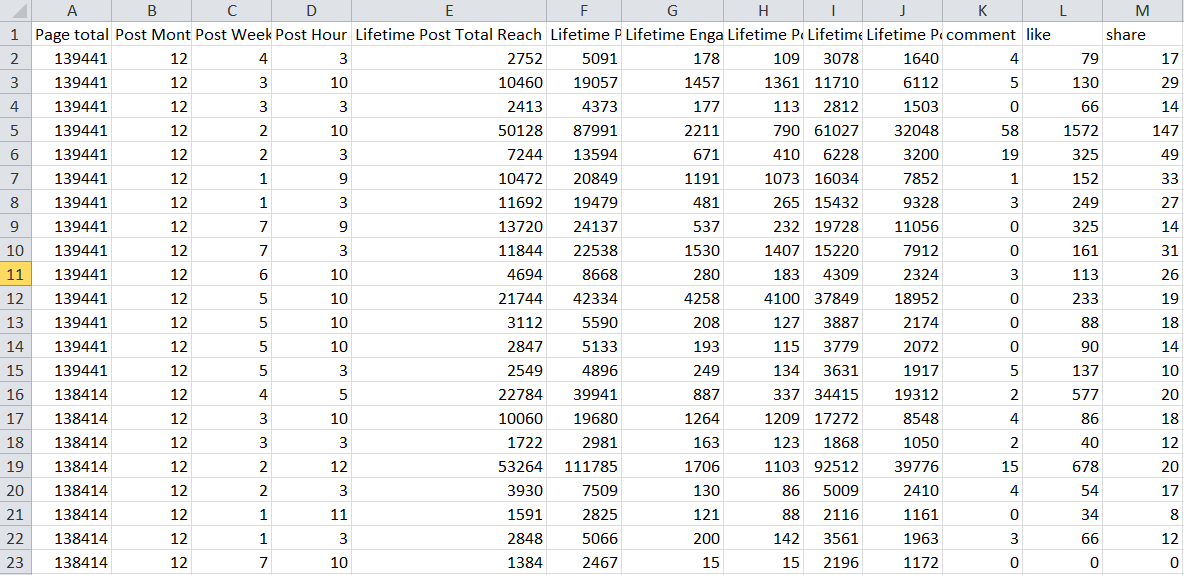
**Интеллектуальные системы и технологии №1**

**Факторный анализ**

|  |  |
| --- | --- |
| *Преподаватель*:  Добренко Н.В | *Выполнил*:  Пыхтин Михаил Андреевич ФИКТ  *Группа К3342* |

Ход работы

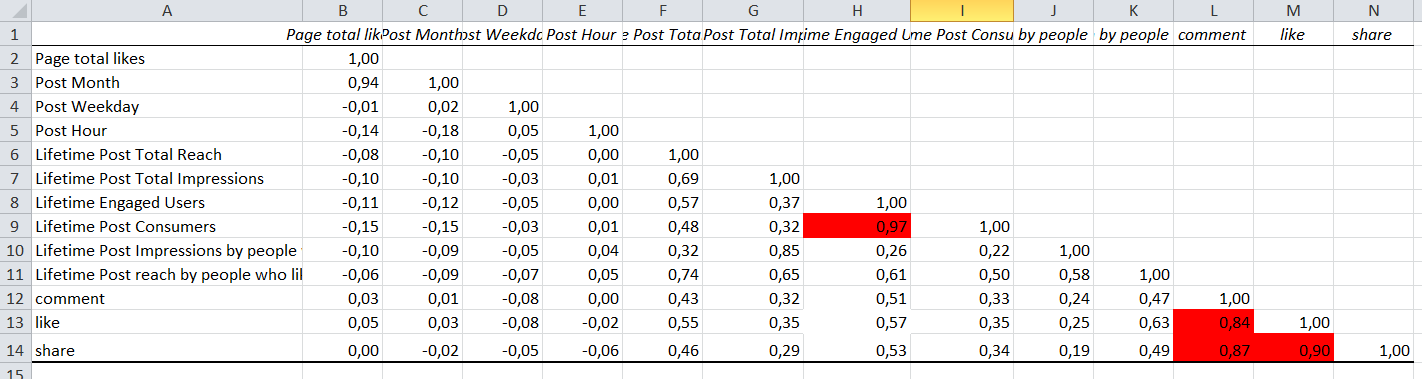
Была произведена выборка данных Фейсбука(http://archive.ics.uci.edu/ml/datasets/Facebook+metrics). И мы выбрали следующие параметры



1. Всего лайков
2. Постов в месяц
3. Постов в неделю
4. Постов в час
5. Количество людей, которые увидели пост
6. общее количество показов
7. общее количество кликов
8. общее количество кликов, из-за лайка страницы
9. количество просмотров
10. количество людей, которые увидели пост, из-за того, что лайкали страницу
11. комментарии
12. лайки
13. поделились

И построим корреляционную матрицу

Которая представляет собой симметричную квадратную матрицу размером *M\*M*, где *М* – число исследуемых факторов, главная диагональ которой заполнена единицами (или нулями для удобства дальнейшего анализа), а недиагональные элементы представляют собой меру тесноты связи между парой факторов (коэффициент корреляции, корреляционное отношение, модифицированный индекс Фехнера и т.д.). Другими словами, для заполнения корреляционной матрицы необходимо найти меру тесноты связи для каждой пары факторов, то есть, провести корреляционный анализ таблицы исходных данных по принципу «каждый с каждым» любым известным способом (например, по методу Чебышева).



# Таким образом была обнаружена статистически достоверная (значимая) связь между изучаемыми переменными share & like, Post Consumers & Engaged users, like & comment. С помощью критических значений корреляции Пирсона

# Затем мы начинаем работу в STATISTICA

Теперь проведем анализ главных компонент и рассмотрим решение с двумя факторами. Для этого рассмотрим корреляции между переменными и двумя факторами, как они были выделены по умолчанию; эти корреляции называются факторными нагрузками.

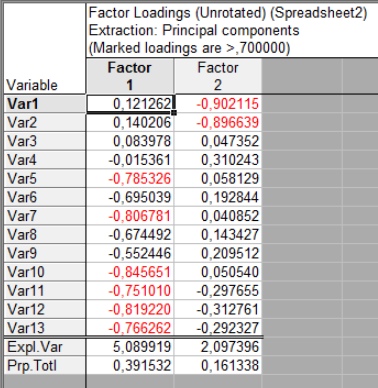
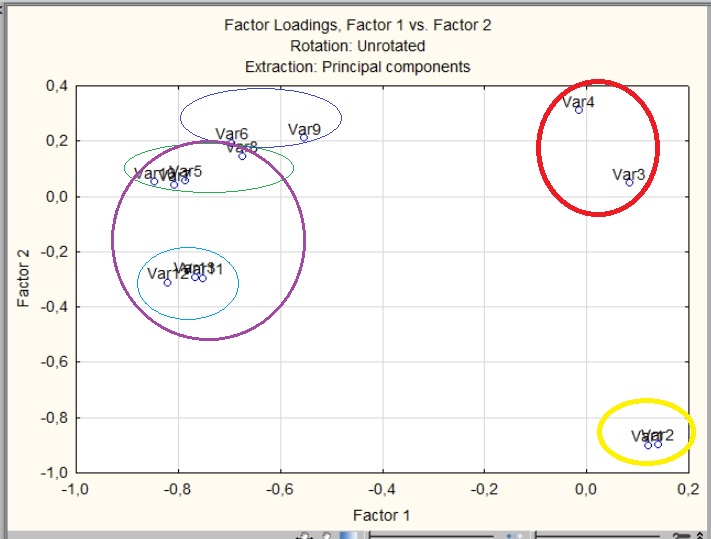


График нагрузки

Фактор 2, как видно из таблицы, имеет наивысшие нагрузки для переменных, относящихся к кол-ву постов.

**Диаграмма рассеивания**



В ней в числах выражена зависимость переменных от факторов. Группируя переменные по наиболее схожим показателям, мы можем понять, какие из них образуют группу на графике.

Таким образом у нас получились группы переменных которые коррелируют друг с другом

**На диаграмме рассеивания факторных нагрузок четко просматривается зависимость:**

Постов в месяц и кол-во лайков

Постов в неделю и постов в час

Количество людей и кол-во людей, подписанных на вас

Клики людей и клики подписчиков

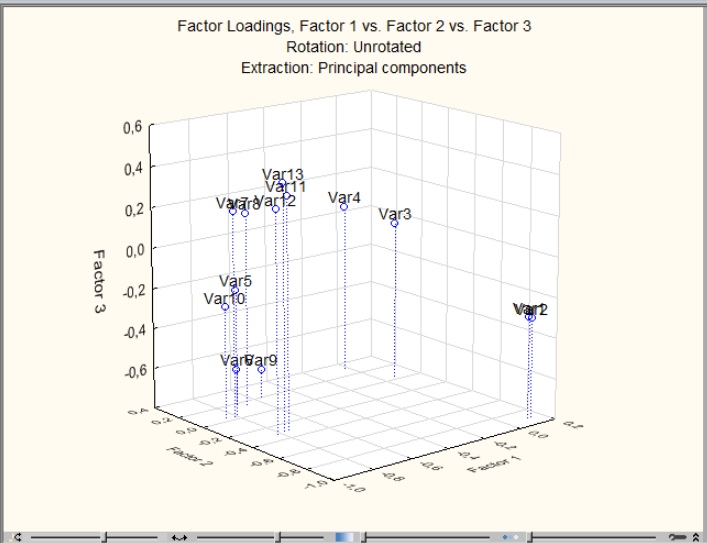
Количество лайков и комментариев

Количество перепостов и лайков

Количество комментариев и перепостов

Также это не противоречит нашей корреляции и здравому смыслу, из чего можно сделать вывод, что факторный анализ выполнен верно.

Для подтверждения посторим так же 3D-график



*Как видите, на данном графике наши зависимости факторного анализа подтвердились. Вся проделанная работа снова доказывает правильность выполненного анализа.*

**Критерий каменистой осыпи.**

# C:\Users\Михаил\Desktop\pMa9y4nvdcY.jpg

является графическим методом, впервые предложенным Кэттелем (Cattell, 1966). Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только "факториальная осыпь" - "осыпь" является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона.

***Различия и сравнения с другой работой***

У моих коллег были отобраны изначально другие признаки, в то время как я сначала использовал факторный анализ на полный датасет, а затем отбросил наименее коррелирующие признаки, и уже проводил факторный анализ на оставшихся. Кроме того, они не использовали критические значения корреляции таблицы Пирсона.

|  |  |
| --- | --- |
| **Page total likes – количество лайков за все время** | Всего лайков |
| **Post month – месяц поста** | Постов в месяц |
| **Post weekday – день недели поста** | Постов в неделю |
| **Post hour – час поста** | Постов в час |
| **Paid – рекламный пост или нет** | Количество людей, которые увидели пост |
| **Unique visits – уникальные посетители** | общее количество показов |
| **People who liked your page – минимальная температура** | общее количество кликов |
| **Comment – комментарии** | общее количество кликов, из-за лайка страницы |
| **Like – лайки** | количество просмотров |
| **Share – репосты** | количество людей, которые увидели пост, из-за того, что лайкали страницу |
|  | комментарии |
|  | лайки |
|  | поделились |

Из-за этих причин, факторный анализ получился точнее у меня впоследствии, что отразилось на дугих графиках уже в программе STATISTICA.

**Вывод:**

*Таким образом, мы выделили наиболее показательных системных счётчиков, которые косвенно могут давать нам информацию об остальных параметрах системы с помощью корреляционных матриц и факторного анализа*