Big Data Applications in Entrepreneurship

In collaboration with Warwick Business School

Dimosthenis Stefanidis
dstefa02@cs.ucy.ac.cy

Nicos Nicolaou
Nicos.Nicolaou@wbs.ac.uk

George Pallis
gpallis@cs.ucy.ac.cy

Marios Dikaiakos
mdd@cs.ucy.ac.cy

Overview

Why Data Analysis Matters in Entrepreneurship?

1. Data helps to uncover hidden patterns in entrepreneur characteristics and behaviors.
2. Can reveal how physical attributes and emotions influence entrepreneurial success (also highlight biases between demographics).
3. Provides actionable insights for investors, researchers, and entrepreneurs.

Outline

1. Existing Theory & Research Questions
2. Data Collection & Preprocessing
3. Mapping Research Questions to Variables
4. Exploratory Data Analysis
5. Model Building/Selection
6. Interpretation of Results
7. Robustness Tests

Existing Theory & Research Questions

Existing Theory / Theoretical Background

- Entrepreneurship is not only influenced by skills but also personal characteristics.

- Facial appearance (e.g., attractiveness) affects perceptions of dominance and competence.

- Emotions, particularly smiling, can impact investor trust and business outcomes.

Facial Appearance

- Scholars have proposed that **stereotyping** and **heuristics** play a **role in social judgments of faces**.

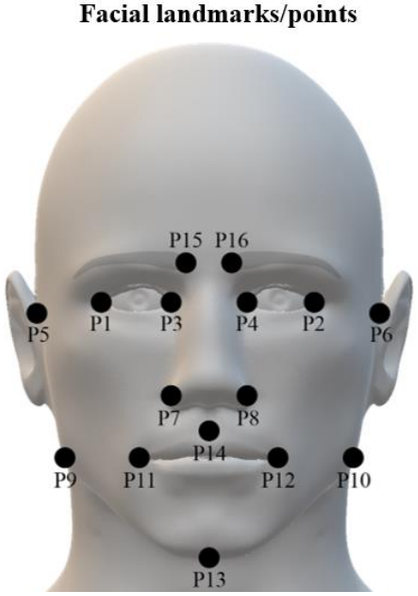
- **Facial appearance** is a basis for forming impressions about a person regarding emotion, personality traits and behavioral intentions. [1][2]

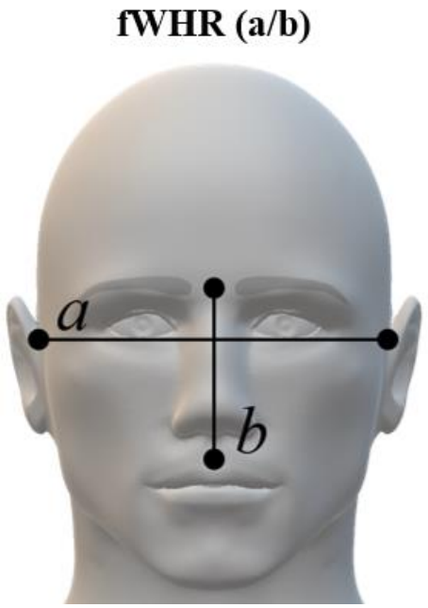
- **Even a few milliseconds exposure** to a face is sufficient to form impressions about a person. [3][4]

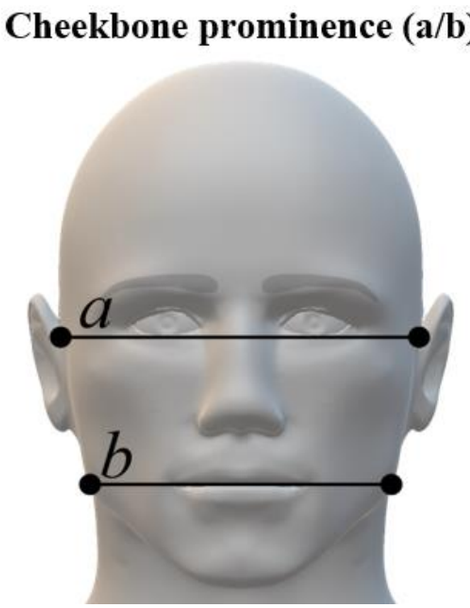
[1] L. A. Zebrowitz, “Finally, Faces Find Favor,” Soc. Cogn., vol. 24, no. 5, pp. 657–701, Oct. 2006.
[2] C. Olivola, F. Funk, A. T.-T. in C. Sciences, and undefined 2014, “Social attributions from faces bias human choices,” Elsevier.
[3] J. Willis, A. T.-P. science, and undefined 2006, “First impressions: Making up your mind after a 100-ms exposure to a face,” journals.sagepub.com.
[4] C. Ballew, A. T.-P. of the National, and undefined 2007, “Predicting political elections from rapid and unreflective face judgments,” Natl. Acad Sci.

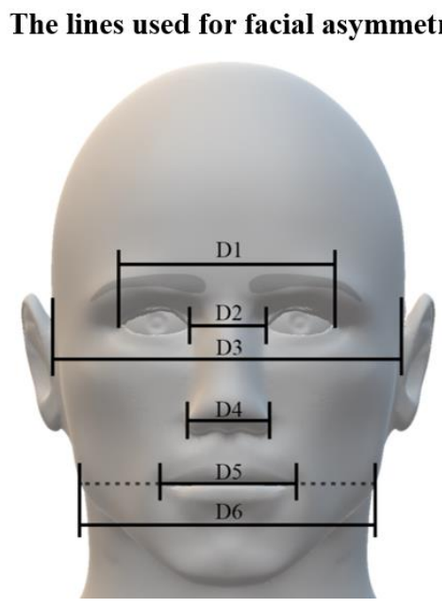
Facial Ratios

- **Facial ratios**
  - 3 facial measures: fWHR, cheekbone prominence and facial symmetry (e.g. overall facial asymmetry and central facial asymmetry)



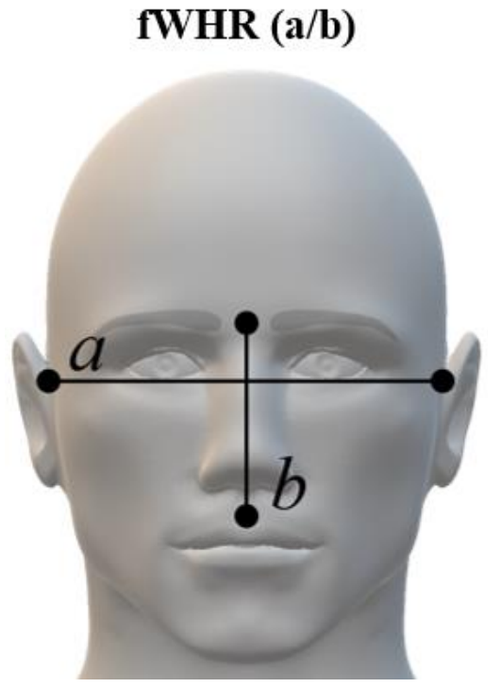






fWHR

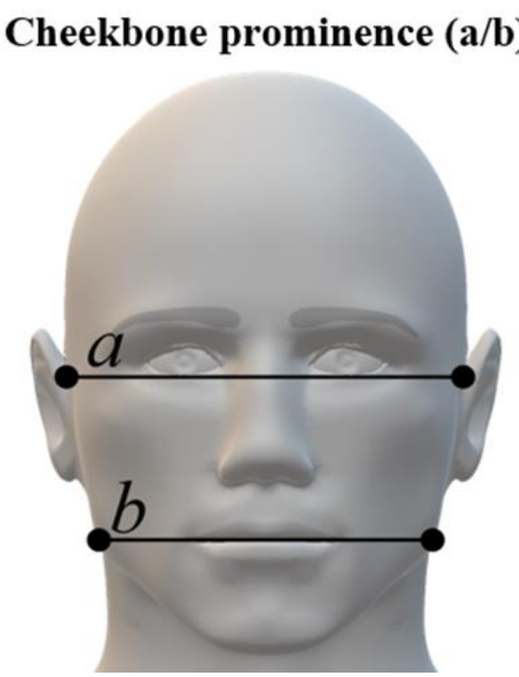
- **Facial width to height ratio (fWHR)** is a measure derived from archaeological skull measurements. [1]
- Faces with **higher fWHR** have been judged as **more competent** and **dominant** [2] and fWHR has been linked with **aggressive behavior** [3] and **risk taking** [4].
- Since **entrepreneurs** are also **associated** with **dominance** [4], **competence** [6] and **risk taking** [5], then people with **higher fWHR** may be more likely to **become entrepreneurs**.



[1] R. S. S. Kramer, A. L. Jones, and R. Ward, “A lack of sexual dimorphism in width-to-height ratio in white European faces using 2D photographs, 3D scans, and anthropometry,” PLoS One, vol. 7, no. 8, p. e42705, 2012.
[2] E. Hehman, J. B. Leitner, M. P. Deegan, and S. L. Gaertner, “Picking teams: When dominant facial structure is preferred,” J. Exp. Soc. Psychol., vol. 59, pp. 51–59, 2015.
[3] J. M. Carré and C. M. McCormick, “In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players,” Proc. R. Soc. London B Biol. Sci., vol. 275, no. 1651, pp. 2651–2656, 2008.
[4] D. L. Sexton and N. Bowman, “The entrepreneur: A capable executive and more,” J. Bus. Ventur., vol. 1, no. 1, pp. 129–140, 1985.
[5] W. H. Stewart Jr and P. L. Roth, “Risk propensity differences between entrepreneurs and managers: A meta-analytic review.,” J. Appl. Psychol., vol. 86, no. 1, p. 145, 2001.
[6] R. A. Baron and G. D. Markman, “Beyond social capital: The role of entrepreneurs’ social competence in their financial success,” J. Bus. Ventur., vol. 18, no. 1, pp. 41–60, 2003.

Cheekbone Prominence

- **Cheekbone prominence** has been **linked** with **attractiveness** [1] [2] and **high masculinity/high testosterone** [3] [4].
- **Attractiveness** has been **associated** with **financial and business success** [5] and **favorable treatment of entrepreneurs** and their ideas [6].
- Since **attractive entrepreneurs** may be able to more **easily draw** the attention of **investors to fund** their ideas into a new successful business, we hypothesize that people with **higher cheekbone prominence** may be more likely to **become entrepreneurs**.



[1] C. F. Keating, A. Mazur, and M. H. Segall, “A cross-cultural exploration of physiognomic traits of dominance and happiness,” Ethol. Sociobiol., vol. 2, no. 1, pp. 41–48, 1981.
[2] M. R. Cunningham, A. P. Barbee, and C. L. Pike, “What do women want? Facialmetric assessment of multiple motives in the perception of male facial physical attractiveness.,” J. Pers. Soc. Psychol., vol. 59, no. 1, p. 61, 1990.
[3] R. E. White, S. Thornhill, and E. Hampson, “Entrepreneurs and evolutionary biology: The relationship between testosterone and new venture creation,” Organ. Behav. Hum. Decis. Process., vol. 100, no. 1, pp. 21–34, 2006.
[4] I. S. Penton-Voak and D. I. Perrett, “Female preference for male faces changes cyclically: Further evidence,” Evol. Hum. Behav., vol. 21, no. 1, pp. 39–48, 2000.
[5] R. Baron and G. Markman, “The role of entrepreneurs’ behavior in their financial success: Evidence for the benefits of effective social skills,” in Babson Conference on Entrepreneurship, Babson Park, MA, USA, 1999.
[6] R. A. Baron, G. D. Markman, and M. Bollinger, “Exporting Social Psychology: Effects of Attractiveness on Perceptions of Entrepreneurs, Their Ideas for New Products, and Their Financial Success 1,” J. Appl. Soc. Psychol., vol. 36, no. 2, pp. 467–492, 2006.

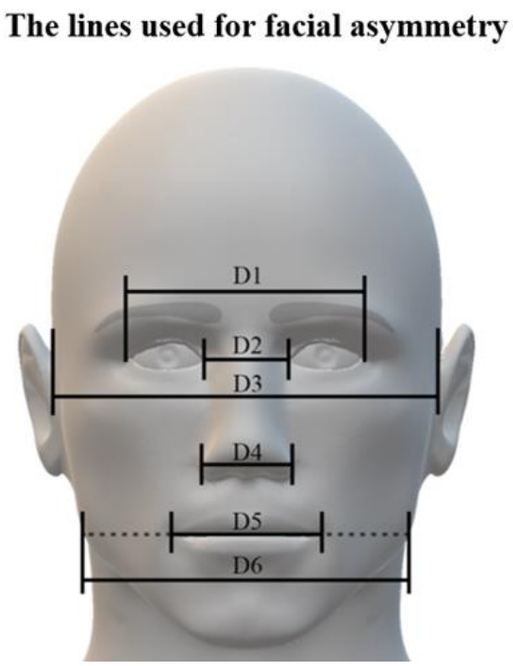
Facial Asymmetry

- **Facial symmetry** (left and right sides of the face resemble) is **associated** with physical **attractiveness** [1] [2] and **health** [3].
- **Facial symmetry** is a possible indicator of **superior genetic quality** ("good genes") and **developmental stability** [4].
- Since entrepreneurship research shows that an **entrepreneur's face** is **associated** with **positive personality behaviors**, **health** and **attractiveness**, we hypothesize that people with **more symmetrical faces** (higher facial symmetry) may be more likely to **become entrepreneurs**.
- **Create Hypotheses using Deductive Reasoning**: A → B, B → C, A → C
  - A = Symmetry, B=Attractiveness, C=Success (DV or y)

[1] K. Grammer and R. Thornhill, "Human (Homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness.," J. Comp. Psychol., vol. 108, no. 3, p. 233, 1994.
[2] D. Jones and K. Hill, "Criteria of facial attractiveness in five populations," Hum. Nat., vol. 4, no. 3, pp. 271–296, 1993.
[3] B. C. Jones, A. C. Little, I. S. Penton-Voak, B. P. Tiddeman, D. M. Burt, and D. I. Perrett, "Facial symmetry and judgements of apparent health: support for a 'good genes' explanation of the attractiveness--symmetry relationship," Evol. Hum. Behav., vol. 22, no. 6, pp. 417–429, 2001.
[4] J. E. Scheib, S. W. Gangestad, and R. Thornhill, "Facial attractiveness, symmetry and cues of good genes," Proc. R. Soc. London B Biol. Sci., vol. 266, no. 1431, pp. 1913–1917, 1999.

Facial Asymmetry

- We focus on **horizontal asymmetry** using the overall facial asymmetry (**OFA**) and the central facial asymmetry (**CFA**) [1].
- **OFA** = the sum of all nonredundant differences between the midpoints of the lines D1, D2, D3, D4, D5 and D6.
- **CFA** = the sum of the differences of the midpoints of the lines Dl and D2, D2 and D3, D3 and D4, D4 and D5, and D5 and D6.
- On a **perfectly symmetrical face**, all **midpoints** must be on the **same vertical line**, and the **OFA** or **CFA** metric is **equal to zero**.



[1] K. Grammer and R. Thornhill, “Human (Homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness.,” J. Comp. Psychol., vol. 108, no. 3, p. 233, 1994.

Facial Emotions

- The face acts as the display board on which **emotions** and **intentions** are **perceived** by **others during social interactions**.

- Different **facial emotions** can **convey** different **mental states** and different **personality traits** (e.g. dominance, affiliation, agreeableness, extraversion, and emotional stability).
  - In business and negotiations settings, smiling conveys friendliness, approval, and satisfaction.

- **Rise of online investment platforms:**
  - It has changed how founders make a first impression on investors.
  - First impressions come from a picture, not in-person.

[1] C. Olivola, F. Funk, A. T.-T. in C. Sciences, and undefined 2014, "Social attributions from faces bias human choices," Elsevier.
[2] P. Borkenau, S. Brecke, C. Möttig, M. P.-J. of R. in, and undefined 2009, "Extraversion is accurately perceived after a 50-ms exposure to a face," Elsevier.
[3] B. Knutson, "Facial expressions of emotion influence interpersonal trait inferences," J. Nonverbal Behav., vol. 20, no. 3, pp. 165–182, 1996.
[4] N. O. Rule and N. Ambady, "The face of success: Inferences from chief executive officers' appearance predict company profits," journals.sagepub.com, vol. 19, no. 2, pp. 109–111, Feb. 2008.
[5] N. O. Rule and N. Ambady, "Face and fortune: Inferences of personality from Managing Partners' faces predict their law firms' financial success," Leadersh. Q., vol. 22, no. 4, pp. 690–696, 2011.
[6] N. O. Rule and N. Ambady, "She's got the look: Inferences from female chief executive officers' faces predict their success," Sex Roles, vol. 61, no. 9–10, pp. 644–652, 2009.

Facial Emotions

“How should they present themselves? Should they smile or steel their faces coolly in their profile photo?”

Founders Raising Money on Seed Invest (Founders of two different ventures observed on Seed Invest at the same time)

Research Questions

RQ1: Does facial structure (e.g., fWHR, cheekbone prominence, facial symmetry) influence entrepreneurship emergence and success?

RQ2: How do facial emotions (e.g., smiling) affect investor decisions and company funding outcomes?

RQ3: Do entrepreneurs express more positive emotions than non-entrepreneurs on social media platforms like Twitter?

Pilot Studies before Main Analysis

- To ensure that smiling does influence funding, we **conducted two pilot studies**

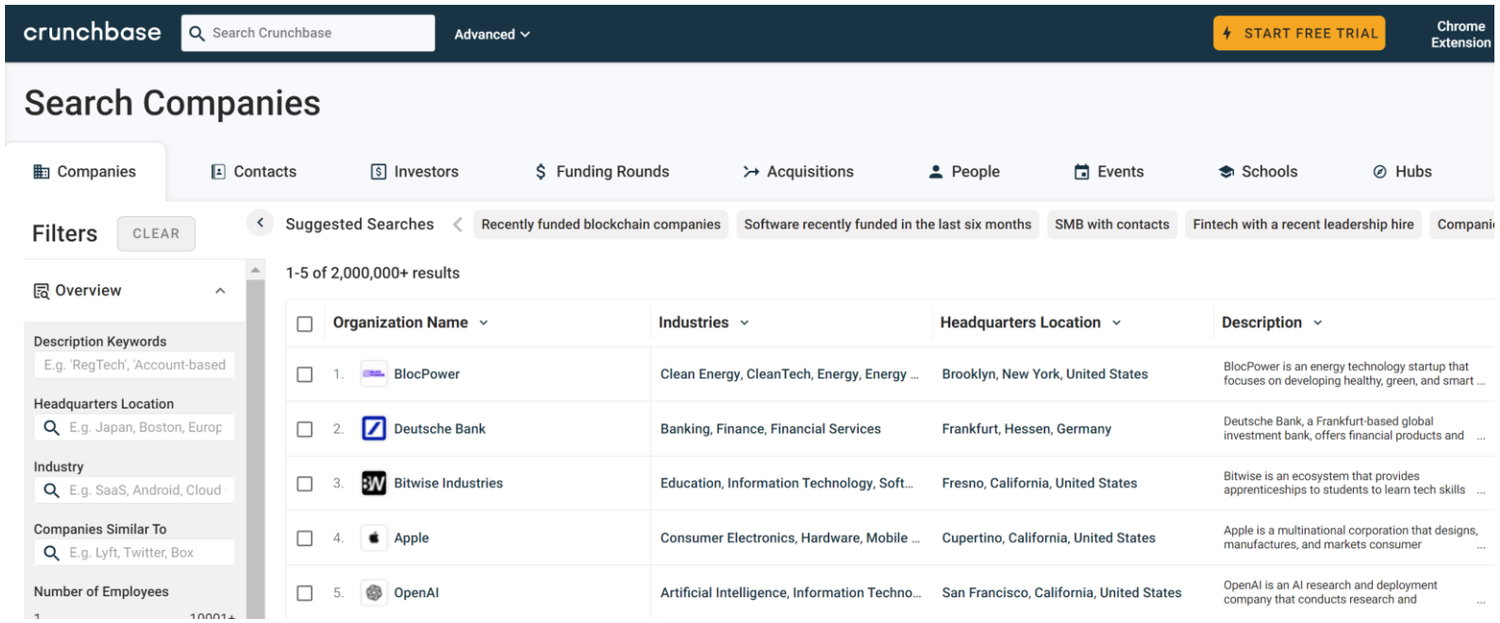
- 1. We **interviewed 10 early stage investors** and asked them whether they favored entrepreneurs who smiled and, if so, why
  - Our findings suggest a **favorable** disposition **towards founders who smile**, **associating** this behavior with **increased trustworthiness**

- 2. We **recruited 210 entrepreneurs** and prospective entrepreneurs to examine the effect of founder smiles on investor interest
  - The results reveal a **significant mediation effect of trustworthiness** between **entrepreneurs' smiling** behavior and investor **fundraising**

Data Collection & Preprocessing/Cleaning

Identify Data Sources

- **CrunchBase** consist of various information about people, companies, funding rounds, investors, products, events, news etc.



crunchbase

URL: www.crunchbase.com

- **Shark Tank,** a popular U.S. pitch competition, where entrepreneu present their ventures to a panel of five investors for potential investment

URL: abc.com/shows/shark-tank

- **Twitter API stream** for consuming tweets real-time.

URL: https://developer.twitter.com/

Tools for Data Collection

- **APIs**: Programmatic interfaces that allow us to access structured data from platforms like:

  CrunchBase API: For company and entrepreneur data.

  Twitter API: For real-time tweet collection.

- **Web Crawlers**: Automated scripts that scrape web pages to gather data.

  Example: Crawling Shark Tank's website to collect pitch data.

  Advantages:

    Speed: Automates data collection.

    Scalability: Collect large volumes of data efficiently.

- **Challenges**:

  Rate limits and restrictions on APIs.

  Cleaning and parsing unstructured data from crawlers.

Additional Tools

- **Face-to-BMI tool** [1] can calculate a person's BMI (body mass index) from profile images by using state-of-the-art computer vision and deep learning techniques.

URL: http://face2bmi.csail.mit.edu/

- **Face++** is a face recognition platform, which capitalizes state-of-the-art deep learning techniques [1][2] and can identify different **face attributes**:
  - gender (male and female)
  - head pose (pitch angle, yaw angle, roll angle)
  - face landmarks
  - emotion recognition (happiness, sadness, surprise, anger, fear, disgust and neutral)

URL: www.faceplusplus.com

[1] H. Fan, Z. Cao, Y. Jiang, Q. Yin, and C. Doudou, "Learning deep face representation," arXiv Prepr. arXiv1403.2802, 2014.
[2] Q. Yin, Z. Cao, Y. Jiang, and H. Fan, "Learning deep face representation." Google Patents, 2016.

Additional Tools

- **uClassify** for identifying the topics of the tweets.

URL: https://www.uclassify.com/

- **Botometer** checks the activity of a Twitter account and gives it a score based on how likely the account is to be a bot. Higher scores are more bot-like.

URL: https://botometer.iuni.iu.edu/

- **NTLK VADER** is a rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

URL: https://github.com/cjhutto/vaderSentiment

Python vs R vs Statistical Software for Data Collection/Preprocessing

Feature	Python	R	Statistical Software (SPSS, STATA)
Flexibility	Highly flexible for web scraping and APIs (e.g., requests, BeautifulSoup, tweepy)	Great for statistical analysis, limited in web scraping	Limited flexibility, mostly GUI-based
Libraries	Rich ecosystem (pandas, numpy, scikit-learn)	Extensive packages for statistics (e.g., dplyr, ggplot2)	Built-in analysis functions, but fewer customization options
Ease of Use	Steep learning curve for non-programmers	Easier for statistics but less suited for complex data manipulation	User-friendly for non-programmers, visual interface
Scalability	Excellent for large datasets, supports distributed computing	Good for mid-sized data but slower on large datasets	Limited, better suited for smaller datasets
Community & Support	Large, active community with lots of online resources	Strong statistical support, large academic community	Limited compared to open-source tools (**also expensive!**)
Best For	Complex data collection, machine learning, and analysis	Data analysis and visualization	Preprocessing and basic statistical analysis

Data preprocessing

Why Data Preprocessing is Important?

- Converts raw data into a suitable format for analysis.
- Enhances model performance by ensuring consistency and reliability.
- Helps in revealing insights that may be obscured in raw data.

Techniques:

- **Encoding:** Convert categorical variables into numerical format (e.g., one-hot encoding).
- **Normalization:** Scale numerical values to a standard range (e.g., min-max scaling).
- **Feature Extraction:** Derive new variables from existing data (e.g., facial ratios, sentiment scores).

Data preprocessing

Encoding Categorical Variables

- **Purpose:** Prepare categorical data (e.g., gender, race) for machine learning models.
- **Method:** One-hot encoding to create binary columns for each category.
- **Example:** Converting the variable "gender" into two columns: "is_male" and "is_female.“

Data preprocessing

Normalization of Numerical Variables

- **Purpose**: Ensure numerical features are on a similar scale to improve model convergence.
- **Method**: Min-max scaling or Z-score normalization.
- **Example**: Scaling age and BMI to a 0-1 range or standardizing to mean 0 and variance 1.

Data preprocessing

Feature Extraction from Facial Images and Text

- **Facial Ratios**: Calculate metrics like fWHR (facial width-to-height ratio) and cheekbone prominence.
- **Sentiment Scores**: Analyze tweet data to derive average happiness or positivity scores using NLTK VADER.
- **Other Features**: Extract face landmarks and emotions using Face++ API.

Data preprocessing

Handling Missing Values

- Univariate vs. Multivariate Imputation
- **Imputation Techniques**: Replace missing values with mean, median, or mode based on the feature type (or even more advance techniques like multiple imputation, KNN, Iterative Imputer).
- **Exclusion Criteria**: Remove entries with excessive missing values to maintain data quality.
- **Importance**: Reduces bias and improves model accuracy.

Data preprocessing

Creating the Final Data Set

- Consolidate transformed data into a single dataset for analysis.
- Ensure consistency across all features (e.g., same units, scales).
- Save the final dataset in a structured format (e.g., CSV, database).

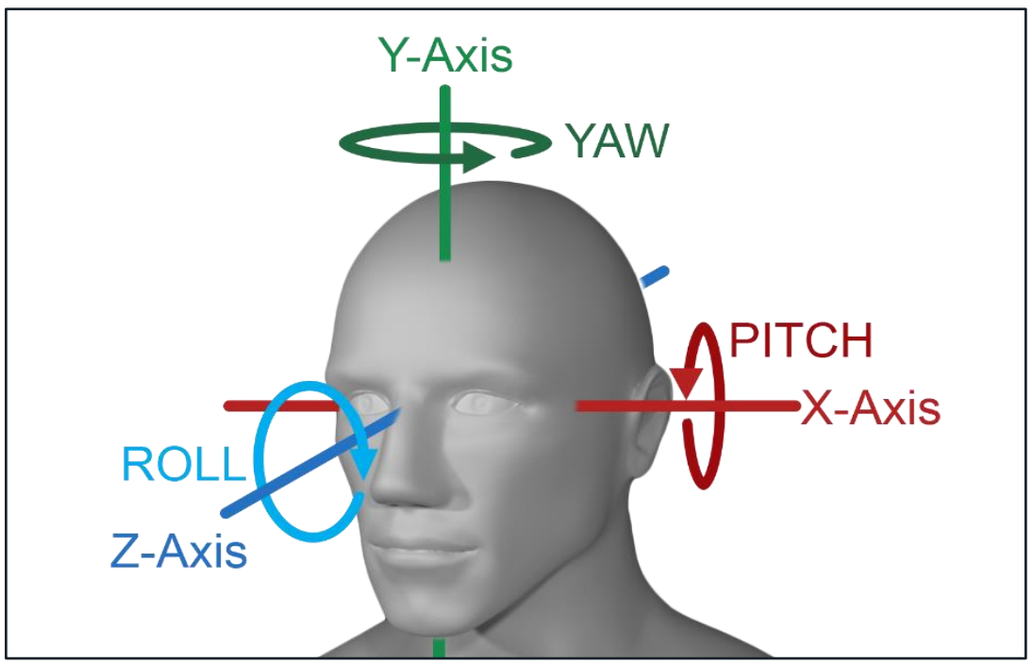
RQ1

- Analysis of the **facial structure** of entrepreneurs and non entrepreneurs
  - **Examine** how **facial ratios** (fWHR, cheekbone prominence, facial symmetry) are **associated** with **entrepreneurship** and **entrepreneurship success**.

- Data collection and preprocessing
  1. **Collect** the profiles of 650.668 **people** from CrunchBase (August 2017)
  2. **Consider** as **entrepreneurs** all the users whose **job title contains** any of the following terms: “entrepreneur”, “founder”, “co-founder”, “owner”, “co-owner” (the remaining users are classified as non-entrepreneurs)
  3. **Keep** all the **users** who have a **profile picture**
  4. **Perform a random sampling** to the **non-entrepreneur** class so as to have a similar size sample with the entrepreneurs’ class
  5. **Use Face++** to **identify** the number of human faces that are depicted in a profile image, along with a series of face-related attributes (e.g. age, gender, race, head pose, face landmarks, emotions)
  6. **Exclude** all the profile pictures that **depict more than one** face

RQ1

7. **Keep** only **“Males”** whose ethnicity is **“White”** and their location is at **“USA”** to **control** for potential **confounders** (e.g. racial differences in facial features)
8. **Exclude** all the users’ profiles that **do not contain information** about the location of their **region** (e.g. California, New York, etc.)
9. **Use Face-to-BMI tool** to estimate the BMI value for each facial photograph
10. **Eliminate** the **outliers** of “Age” and “BMI”
11. **Keep** only the images where **face rotation (pitch, yaw and roll**) is **within a specific range** as it affects the precision of the detection of facial landmarks



RQ1

Steps	Entrepreneurs	Non-Entrepreneurs
Download the profile of people in CrunchBase in August 2017	152900	497768
Users who have a profile picture in CrunchBase	101771	336589
Random sampling of non-entrepreneurs	101771	110000
One face depicted in the profile image	92216	98339
Male	80526	78040
White	64427	66790
USA	21764	24840
Profiles that contain region information	21719	24774
Age ≥ 18	21709	24765
18.4 ≤ BMI ≤ 40.5	21289	24255
-30o ≤ Yaw ≤ 30o and -30o ≤ Pitch ≤ 30o and -30o ≤ Roll ≤ 30o	20771	23941
-10o ≤ Yaw ≤ 10o and -10o ≤ Pitch ≤ 10o and -10o ≤ Roll ≤ 10o	11679	14464
-5o ≤ Yaw ≤ 5o and -5o ≤ Pitch ≤ 5o and -5o ≤ Roll ≤ 5o	3032	3787

RQ2

- Analysis of the **facial emotions** of entrepreneurs and how these emotions affect the **financial success** of their company
  - Examine whether **founders who smile** are more likely to **influence** the **investors** decisions and **receive** a **funding.**

- Data collection and preprocessing
  1. **Collect** 528059 **companies**' profiles, 527663 **founders**' profiles and 178416 **funding rounds** information from CrunchBase (October 2017)
  2. **Keep** all the **founders** who have a **profile picture**
  3. **Use Face++** to **identify** the number of human faces that are depicted in a profile image, along with the facial expressions/emotions of the faces
  4. **Exclude** all the profile pictures that **depict more than one** face
  5. **Keep** the companies with **headquarters location** at **"USA" and** their **founded year** is within the range **[2011, 2015]** to **control** for potential **confounders** (e.g. regional influences, microeconomic factors)

RQ2

Preprocessing Steps	Observations of the Dependent Variable		
	Number of rounds (Negative Binomial regression)	Has receive funding (Logistic regression)	Funding received (Tobit, Cragg Hurdle)
(1) Download companies' profiles, founders' profiles and funding rounds info from CrunchBase in October 2017	528051	528051	494880
(2) At least one of the founders had a profile image that depicted only one face	67281	67281	56345
(3) USA	33597	33597	27455
(4) Founded Year [2011, 2015]	20316	20316	16135

RQ3

- Examine whether **entrepreneurship influences an individuals' emotions** using more than 30 million Twitter messages sent by entrepreneurs and non-entrepreneurs.

- Data collection and preprocessing

1. **Collect** tweets and user's profiles from the **streaming API of Twitter** (January 2013 to January 2015)
2. **Identify entrepreneurs** based on **keywords** ("entrepreneur", "founder", "co-founder", "business-owner", "business owner", "start-up", or "start up") in their personal Twitter description.
3. **Identify non-entrepreneurs** that are **not bots** or **company profiles.**
4. **Identify social entrepreneurs** who describe themselves as such in their Twitter profile, using the keywords "social" and "entrepreneur".
5. **Identify serial entrepreneurs** who describe themselves as such in their Twitter profile, using the keywords "serial" and "entrepreneur".
6. **Use NLTK Vader** to **identify** the **sentiment** of each **tweet.**
7. **Use uClassify** to **identify** the **topic** of each **tweet.**

RQ3

Table: Number of users, initial tweets, and usable tweets for the **main results**

Preprocessing Steps	**Raw WW Data** from January 2013 to January 2015 Users: 37.390.316 Tweets: 13.527.604.878
(1) Find Entrepreneurs (based on keywords in the description)	Entre: 24.573 Entre Tweets: 39.641.296
(2) Find 24.573 non-entrepreneurs that are not bots* or company profiles**	Non-Entre: 24.573 Non-Entre Tweets: 22.426.213

*Detect bots using Botometer tool (Ferrara et al, 2016; Davis et al, 2016; Varol et al, 2017)

**Companies profile have far more followers than followings (Chu et al, 2010)

RQ3

Table: Number of users, initial tweets, and usable tweets for extracting the **discussion topic**

Preprocessing Steps	**WW Data** from September 2014 to November 2014 Tweets: 3.184.087
(1) Remove tweets that contain only urls, user mentions or symbols	Tweets: 3.071.587
(2) Remove tweets with low class/topic probability score (the highest class probability must be bigger than 0.9)	Tweets: 304.323

Mapping Research Questions to Variables

Defining Dependent, Independent, and Control Variables

Understanding Variables in Research

- **Dependent Variables (DV) (also known as target/outcome variables):** The outcomes or responses that researchers are interested in explaining or predicting. They are influenced by the independent variables in the study.

- **Example:** The variable isEntrepreneur (1 for entrepreneurs, 0 for non-entrepreneurs) serves as a dependent variable because you are trying to determine how various factors (like facial features and emotions) influence a person's status as an entrepreneur. Other examples of DVs include Revenue and Valuation, which reflect the success of the entrepreneurial venture.

Understanding Variables in Research

- **Independent Variables (IV) (also known as features):**

  The predictors or factors that are hypothesized to influence the dependent variables. Researchers manipulate or categorize these variables to see how they affect the DVs.

- **Example:** Variables such as fWHR, Cheekbone Prominence, and Facial Symmetry are independent variables in your analysis. You are examining how variations in these facial characteristics impact the likelihood of someone being an entrepreneur or the success of their business.

Understanding Variables in Research

- **Control Variables:** Control variables are additional factors that researchers account for to minimize their influence on the dependent variable. By holding these variables constant, researchers can isolate the effects of the independent variables more accurately.

- **Example**: Variables like Age, Gender, and Location are crucial control variables in your research. For instance, age can influence entrepreneurial success due to experience, while gender may affect perceptions of leadership and competence. By controlling for these factors, your analysis can more reliably attribute changes in the dependent variables to the independent variables being studied.

RQ1 - Variables

- **Dependent Variable**
  - **isEntrepreneur:** "1" for entrepreneurs and "0" for non-entrepreneurs (binary)
  - **Revenue:** the revenue of a company (continuous)
  - **Receive funding**: whether a company has received at least one funding round or not, with values 1 and 0 respectively (binary)
  - **Total funding amount:** the total amount that was raised across all funding rounds (continuous)
  - **Valuation:** the most recent valuation of a company (continuous)
- **Independent Variables**
  - fWHR
  - Cheekbone prominence
  - Facial Symmetry

RQ1 - Variables

- **Control Variables**
  - **Region/Location**: the entrepreneurial activity is different in each geographical location
  - **Age**: older individuals may be more likely to be entrepreneurs
  - **Rotation angle of a face (yaw, pitch and roll):** rotation may affect the facial ratios
  - **BMI:** is linked with fWHR and cheekbone prominence and may affect them
  - **Education:** Education may affect an individual's decision to start a business
  - **Industry:** Different industries are associated with different rates of performance and success in attracting funding

RQ2 - Variables

- **Dependent Variables**
  - **Receive funding**: whether a company has received at least one funding round or not, with values 1 and 0 respectively
  - **Number of rounds:** the "total number of funding rounds" that a company has received
  - **Total funding amount:** the total amount that was raised across all funding rounds

- **Independent Variable**
  - **“Average happiness”:** the average happiness of the founders of a company that was calculated based on the happiness of each founder’s image

$$\frac{founder\_1\_happiness\_confidence + \cdots + founder\_N\_happiness\_confidence}{N}$$

N = number of founders for a company

RQ2 - Variables

- **Regression Analysis**
  - **Number of rounds (≥0):** Negative Binomial Regression Model (due to overdispersion)
  - **Receive funding (0/1)**: Binary Logistic Regression Model
  - **Total funding amount (≥0):** Tobit Regression Model and Cragg Hurdle Regression Model (far more zeros observed than expected)
  - **Transformed total funding amount (>0):** Linear Regression Model

RQ3 - Variables

- **Dependent Variables**
  - **Sentiment:** average sentiment of each user based on all of their published tweets.

- **Independent Variables**
  - **isEntrepreneur:** "1" for entrepreneurs and "0" for non-entrepreneurs
  - **is Social Entrepreneur:** "1" for social entrepreneurs and "0" for other entrepreneurs
  - **is Serial Entrepreneur:** "1" for serial entrepreneurs and "0" for other entrepreneurs

Exploratory Data Analysis

Purpose of EDA

Why Perform Exploratory Data Analysis?

- Identify trends, patterns, and anomalies within the data.
- Understand the distribution and relationships between variables.
- Inform decisions on subsequent data analysis methods (e.g., modeling techniques).
- Validate Research Questions/Hypotheses

Key Techniques in EDA

- **Descriptive Statistics**: Summarize data characteristics using measures like mean, median, mode, variance, and standard deviation.

- **Data Visualization**: Use graphs and plots to visualize data distributions and relationships (e.g., box plots, histograms, scatter plots).

- **Correlation Analysis**: Examine relationships between variables to identify potential associations (e.g., correlation coefficients).

- **Outlier Detection**: Identify and assess the impact of outliers on the dataset.

Descriptive Statistics

- **Purpose:** Provide a summary of the main characteristics of the data.

- **Key Measures:**

  Mean: Average value of a variable.

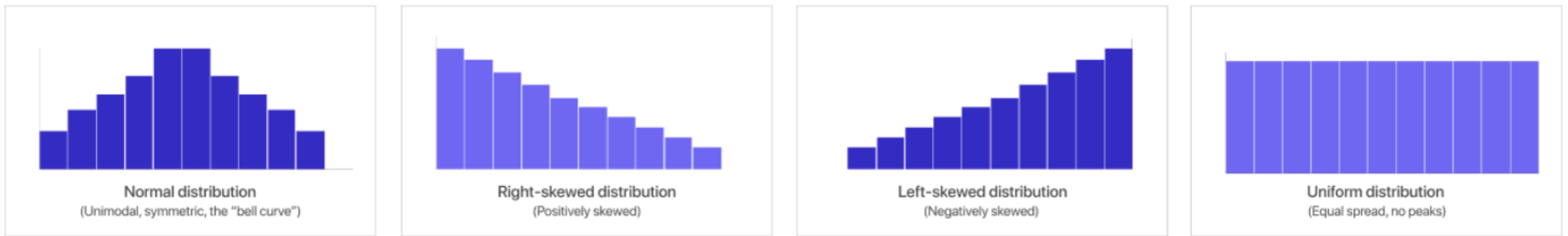
  Median: Middle value when the data is ordered.

  Standard Deviation: Measure of variability or dispersion in the dataset.

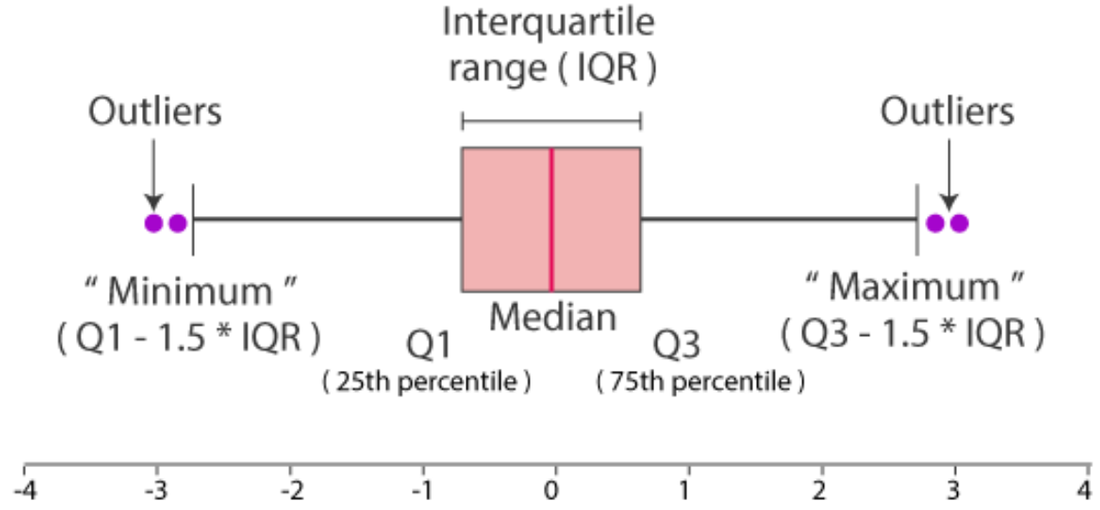
  Counts: Number of observations for categorical variables (e.g., count of entrepreneurs vs. non-entrepreneurs).

Data Visualization

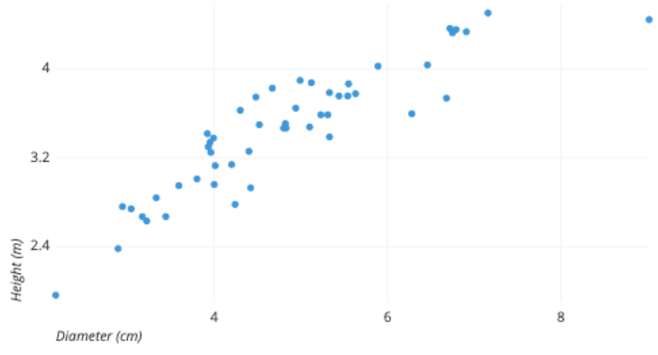
- **Histograms:** Show the distribution of continuous variables (e.g., age, revenue).



- **Box Plots:** Visualize the distribution and identify outliers in variables (e.g., funding amounts).



- **Scatter Plots:** Explore relationships between two numerical variables (e.g., facial symmetry vs. revenue).



Correlation Analysis

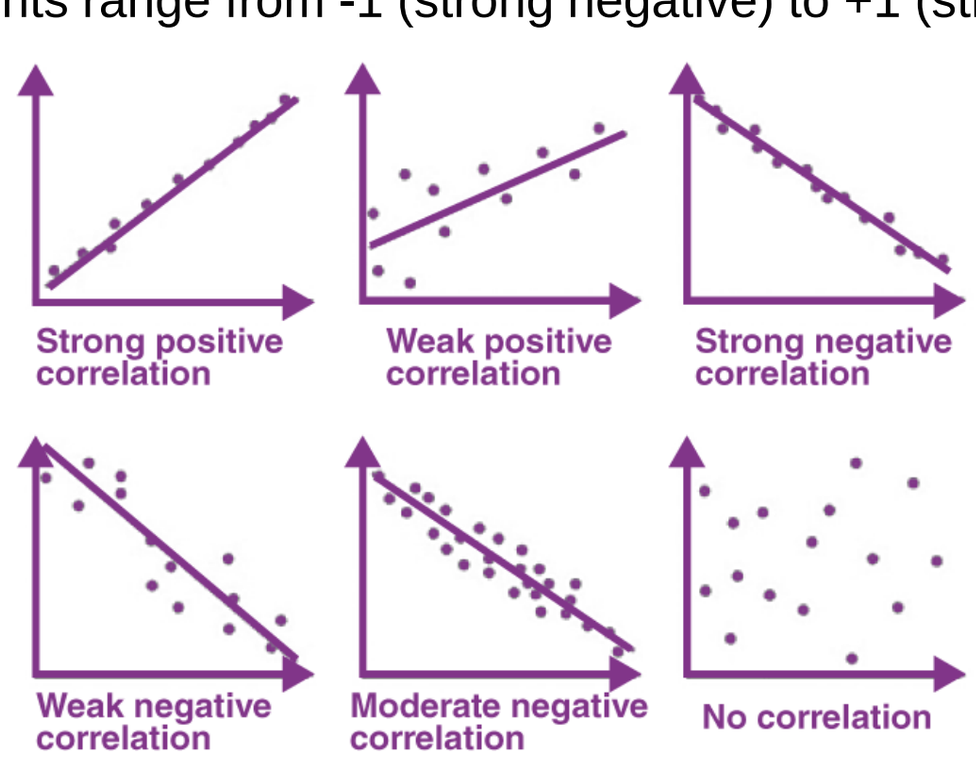
Understanding Variable Relationships

- **Purpose:** Determine the strength and direction of relationships between variables.
- **Methods – e.g.:**

  Pearson Correlation Coefficient: Measures linear correlation between two variables.

  Heatmaps: Visual representation of correlation coefficients between multiple variables.

  Interpretation: Coefficients range from -1 (strong negative) to +1 (strong positive), with 0 indicating no correlation.



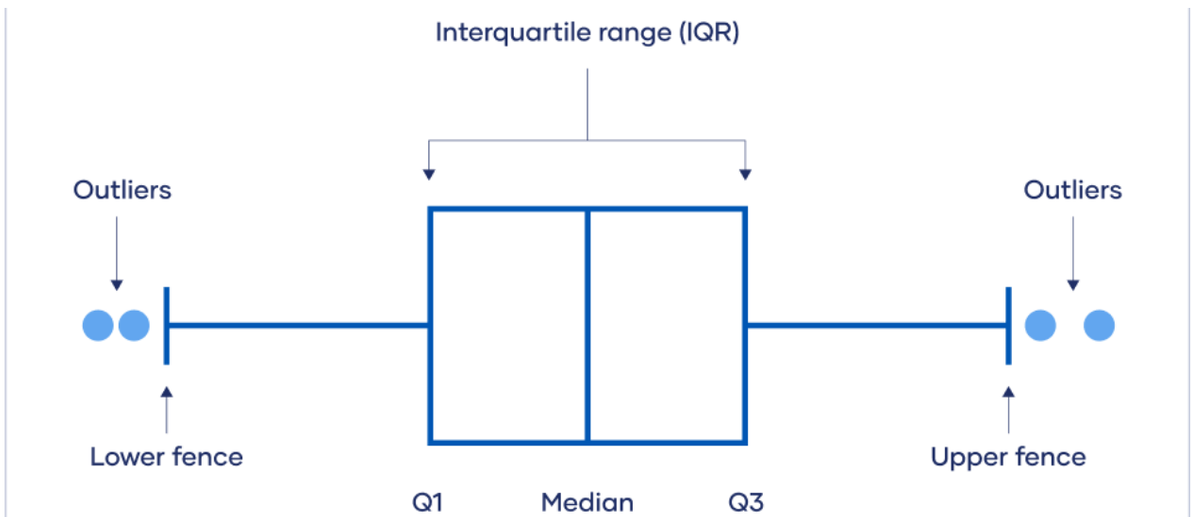
Outlier Detection

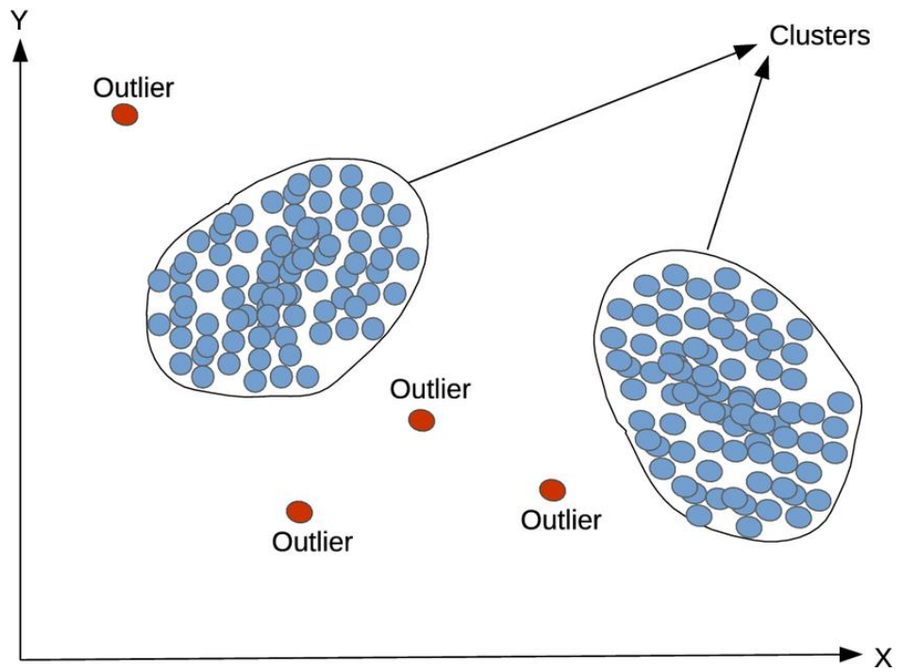
- **Definition:** Outliers are data points that deviate significantly from the rest of the dataset.
- **Techniques:**

  Z-score Method: Identifying outliers based on standard deviations from the mean.

  Box Plot Method: Using the interquartile range (IQR) to determine upper and lower bounds for detecting outliers.

Impact of Outliers: Assess how outliers may affect analysis results and model performance.





Descriptive statistics and correlations

- **Correlation**: quantifies the strength of the linear relationship between a pair of variables
- **The key difference** between **correlation** and **regression** is that correlation measures the degree of a relationship between two independent variables (x and y). In contrast, **regression measures how one variable affects another**.

```
.
.     tabstat Age Roll Yaw Pitch BMI Created_cb_year Bachelor Master PhD fWHR Cheekbone OFA CFA isEntrepreneur, s(mean sd co
> unt) format(%9.3fc)

   stats |       OFA       CFA  isEntr~r
---------+------------------------------
    mean |     0.238     0.076     0.445
      sd |     0.154     0.058     0.497
       N | 6,819.000 6,819.000 6,819.000
------------------------------------------
```

```
             |   Master      PhD     fWHR Cheekb~e      OFA      CFA isEntr~r
-------------+------------------------------------------------------------
             |
        fWHR |   0.0251  -0.0195   1.0000
             |   0.0384   0.1075
             |     6819     6819     6819
             |
   Cheekbone |  -0.0201  -0.0006  -0.1090   1.0000
             |   0.0962   0.9576   0.0000
             |     6819     6819     6819     6819
             |
         OFA |   0.0347   0.0093   0.0004   0.0048   1.0000
             |   0.0042   0.4442   0.9764   0.6924
             |     6819     6819     6819     6819     6819
             |
         CFA |   0.0366   0.0011   0.0282  -0.0081   0.9386   1.0000
             |   0.0025   0.9282   0.0200   0.5034   0.0000
             |     6819     6819     6819     6819     6819     6819
             |
isEntrepre~r |  -0.0179   0.0506   0.0020   0.0674  -0.0654  -0.0711   1.0000
             |   0.1405   0.0000   0.8692   0.0000   0.0000   0.0000
             |     6819     6819     6819     6819     6819     6819     6819
             |
```

Descriptive statistics and correlations

Descriptive statistics and variable correlations.

Variable	Mean	SD	1	2	3	4	5	6	7	8	9	10	11
1. Age[(a)]	42.241	9.433											
2. BMI[(a)]	29.743	3.678	0.165***										
3. Created CB Year[(a)]	2,014.372	2.238	0.027*	0.011									
4. Bachelor[(a)]	0.208	0.406	-0.040***	-0.014	-0.100***								
5. Master[(a)]	0.159	0.365	0.029*	-0.010	-0.145***	-0.222***							
6. PhD[(a)]	0.061	0.239	0.034**	-0.007	-0.093***	-0.130***	-0.111***						
7. fWHR[(a)]	1.891	0.132	-0.023	0.210***	0.010	-0.003	0.025*	-0.019					
8. Cheekbone[(a)]	1.101	0.024	-0.098***	-0.187***	-0.010	0.002	-0.020	-0.001	-0.109***				
9. OFA[(a)]	0.238	0.154	0.143***	0.062***	0.013	-0.025*	0.035**	0.009	0.000	0.005			
10. CFA[(a)]	0.076	0.058	0.139***	0.070***	0.019	-0.023	0.037**	0.001	0.028*	-0.008	0.939***		
11. Entrepreneur[(a)]	0.445	0.497	-0.196***	-0.050***	-0.222***	0.052***	-0.018	0.051***	0.002	0.067***	-0.065***	-0.071***	
12. Revenue[(b)]	8.98e + 07	8.70e + 08	0.044	-0.012	-0.117***	-0.020	0.105***	-0.003	-0.033	-0.006	0.010	0.009	c*

[a] Observations = 6,819, [b] Observations = 1,569

c* The correlation between “Entrepreneur” and “Revenue” cannot be computed because all of the observations in the variable “Revenue” are entrepreneurs. For those observations, the value of “Entrepreneur” is always 1.

*** $p < .001$, ** $p < .01$; * $p < .05$

Analyze Sentiment

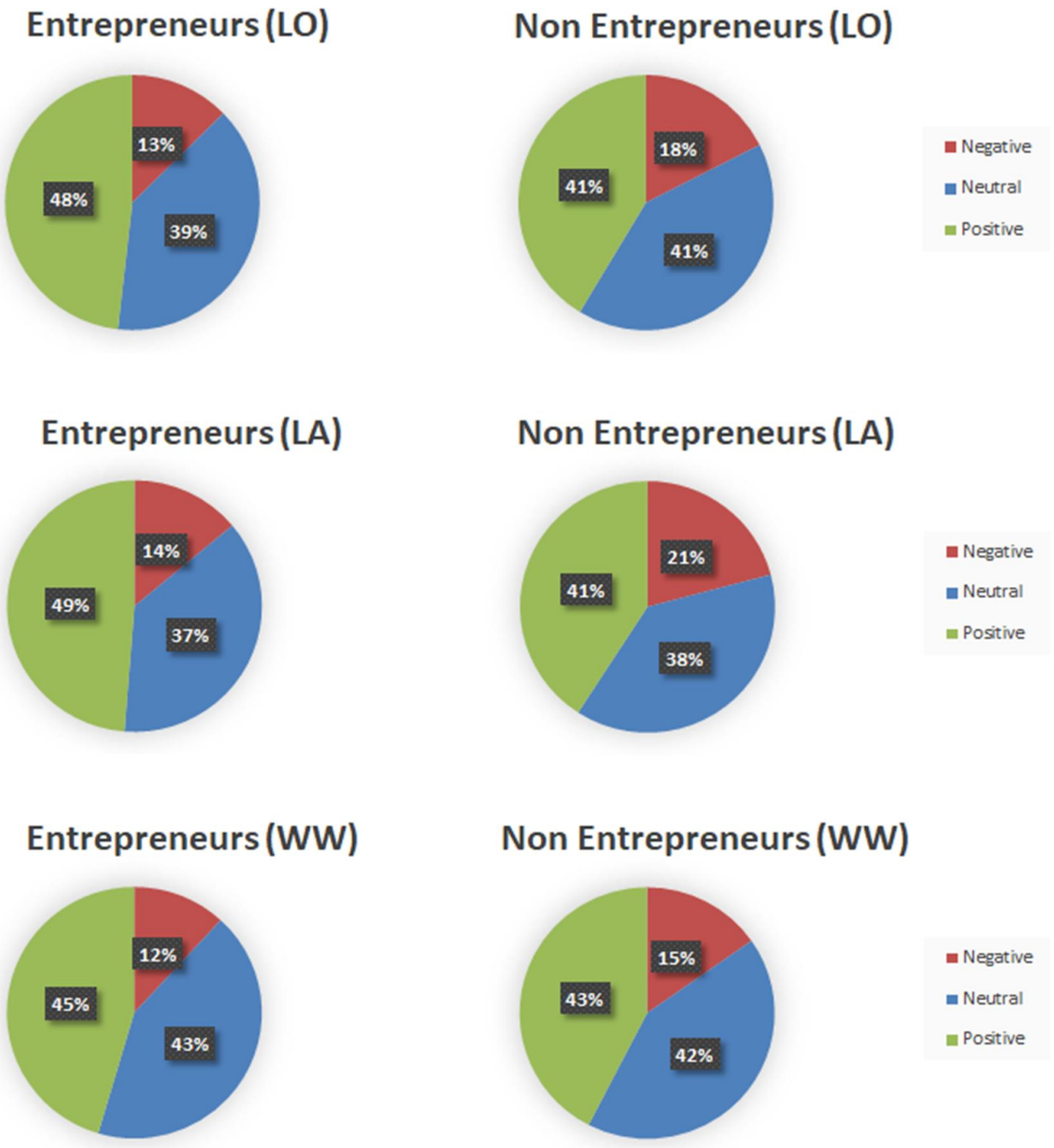
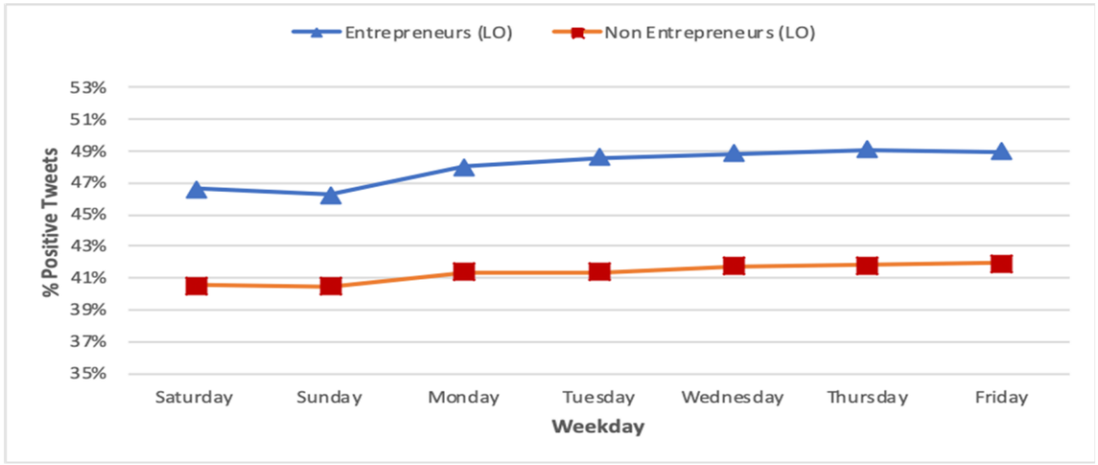
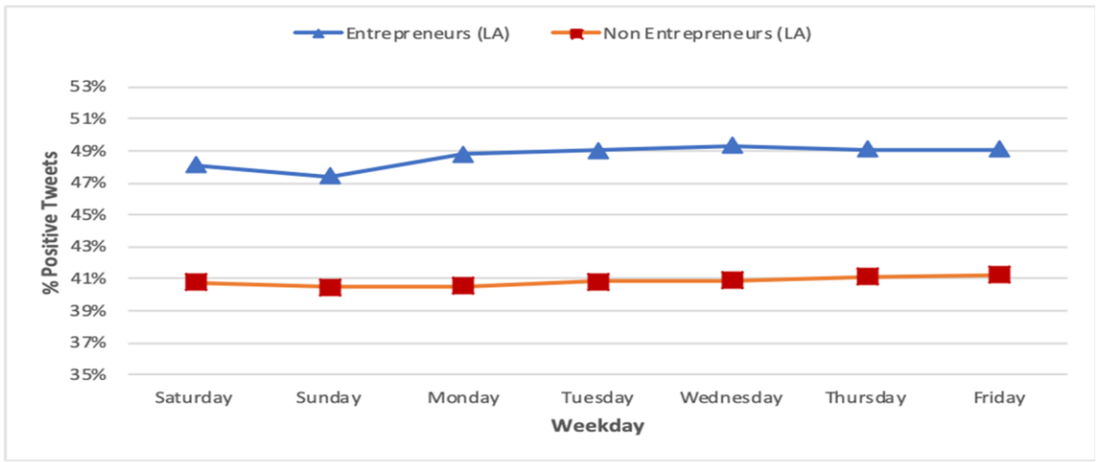


Figure: Tweet emotional comparison between entrepreneurs and non-entrepreneurs.

Analyze Sentiment





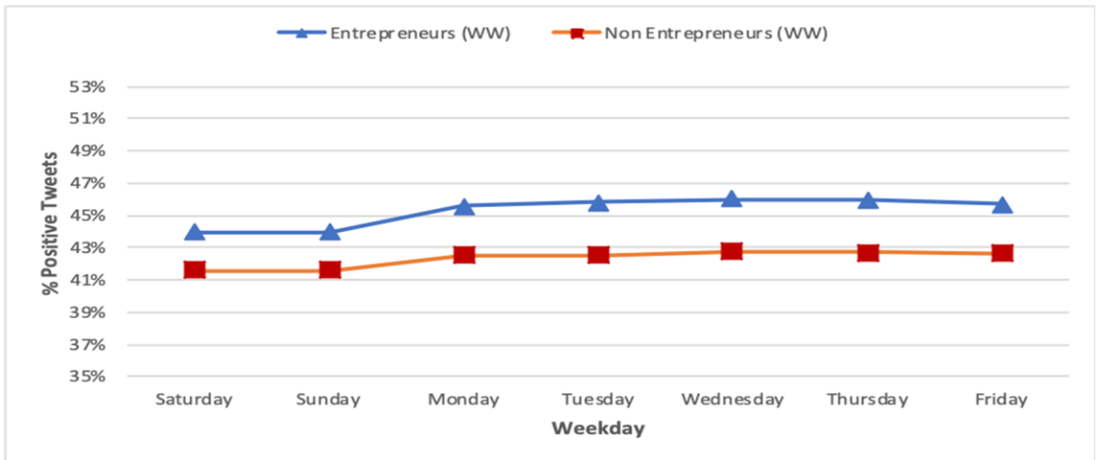
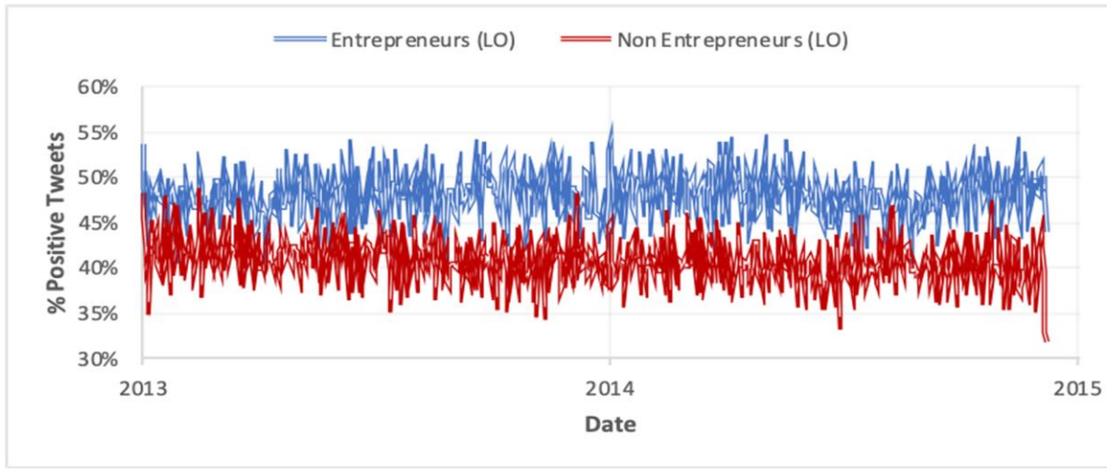
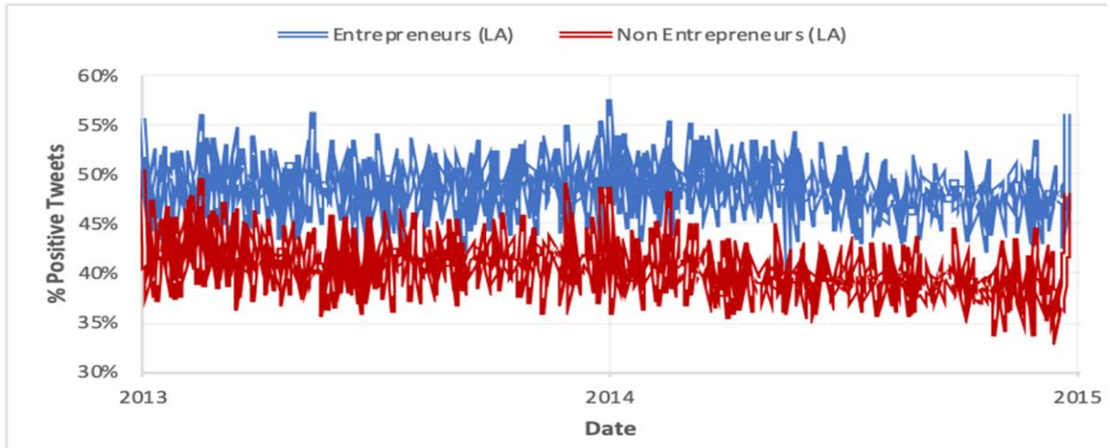


Figure: Comparison between entrepreneurs and non-entrepreneurs on their tweets; emotional score per weekday.

Analyze Sentiment





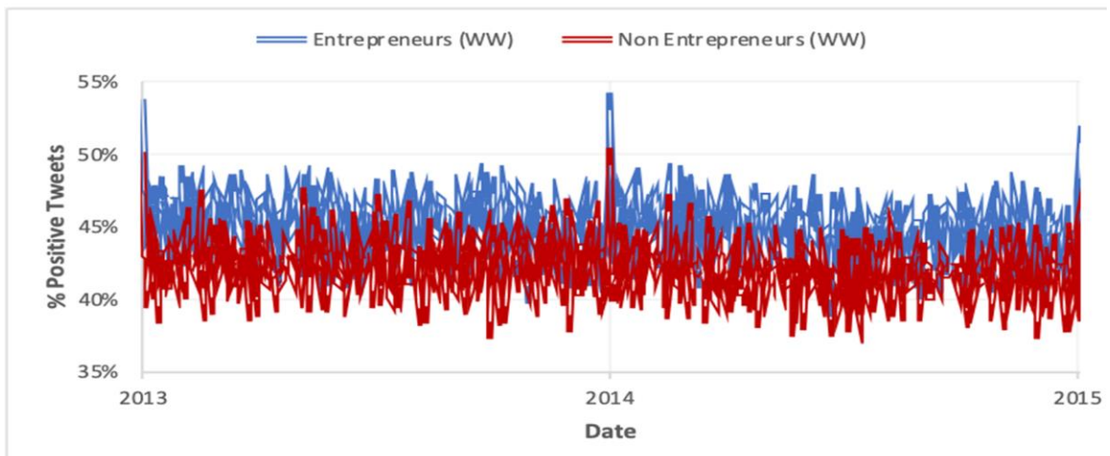


Figure: Emotional score per calendar day, from January 2013 to January 2015, averaged over active users per day.

Analyze Topics

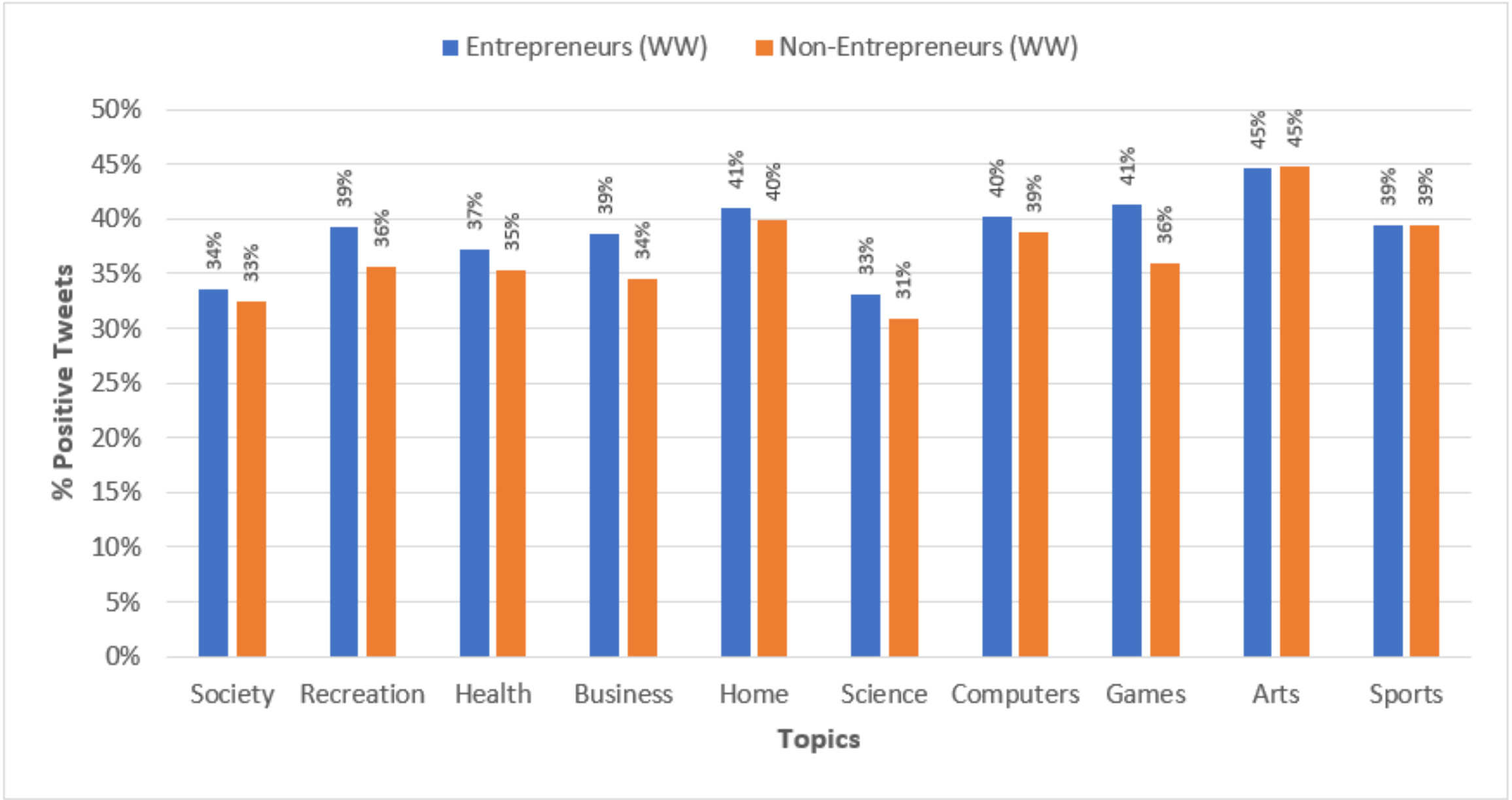


Figure: Emotional score per concept, for a sample of Tweets

Model Building/Selection

Purpose of Model Building

Why Model Building is Important

- To establish relationships between independent and dependent variables.
- To make predictions or infer conclusions based on data.
- To validate research hypotheses using statistical methods.
- To provide actionable insights for stakeholders.

Purpose of Model Building

Types of Models for Analysis (examples)

- **Linear Regression**: Used for predicting continuous outcomes (e.g., revenue based on facial metrics).
- **Logistic Regression**: Suitable for binary outcomes (e.g., predicting isEntrepreneur).
- **Fixed Effects Models**: Used for panel data analysis to control for unobserved variables that do not change over time.
- **Cragg Hurdle Models**: Suitable for count data with excess zeros, allowing separate modeling for zero and positive counts.

Model Selection Criteria

Performance Metrics (examples):

- **R-squared:** Measures the proportion of variance in the dependent variable explained by the independent variables. A higher R^2 indicates a better fit.
- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of the errors in a set of predictions, with lower values indicating a better fit.
- **Mean Absolute Error (MAE)**: Represents the average absolute difference between predicted values and actual values, providing a clear interpretation of average error.
- **Mean Squared Error (MSE)**: Similar to RMSE but squares the errors before averaging, penalizing larger errors more than smaller ones.
- **Accuracy, F1, Precision, Recall:** For classification models to evaluate performance.

Interpretation of Results

RQ1 - Emergence

- **Dependent Variable**
  - **isEntrepreneur:** "1" for entrepreneurs and "0" for non-entrepreneurs (binary)
- **Independent Variables**
  - Facial Symmetry

```
. 
. reg isEntrepreneur ity_0_newyork-City_9_denver Age Created_cb_year Roll Yaw Pitch BMI Bachelor Master PhD OFA, cformat(%9.3f)

Linear regression                               Number of obs     =      6,819
                                                F(20, 6798)       =      41.42
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0973
                                                Root MSE          =     .47286

------------------------------------------------------------------------------
             |               Robust
isEntrepreneur |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 City_0_newyork |     -0.077      0.018    -4.34   0.000       -0.112      -0.042
  City_1_boston |     -0.096      0.026    -3.70   0.000       -0.146      -0.045
City_2_losangeles |      0.055      0.030     1.83   0.068       -0.004       0.114
 City_3_seattle |      0.027      0.037     0.75   0.455       -0.044       0.099
City_4_washington |     -0.118      0.031    -3.77   0.000       -0.179      -0.056
 City_5_chicago |     -0.065      0.028    -2.32   0.020       -0.120      -0.010
  City_6_austin |      0.058      0.041     1.44   0.151       -0.021       0.138
 City_7_sandiego |      0.015      0.051     0.30   0.764       -0.085       0.115
 City_8_atlanta |     -0.021      0.045    -0.48   0.630       -0.109       0.066
  City_9_denver |     -0.022      0.041    -0.54   0.589       -0.101       0.058
            Age |     -0.010      0.001   -15.20   0.000       -0.011      -0.008
 Created_cb_year |     -0.048      0.003   -18.40   0.000       -0.053      -0.043
           Roll |      0.002      0.002     0.81   0.421       -0.003       0.006
            Yaw |      0.001      0.002     0.46   0.647       -0.003       0.005
          Pitch |     -0.003      0.002    -1.21   0.227       -0.007       0.002
            BMI |     -0.002      0.002    -1.56   0.118       -0.006       0.001
       Bachelor |      0.024      0.015     1.66   0.098       -0.004       0.053
         Master |     -0.043      0.016    -2.63   0.009       -0.076      -0.011
            PhD |      0.071      0.025     2.83   0.005        0.022       0.120
            OFA |     -0.102      0.039    -2.61   0.009       -0.179      -0.025
          _cons |     97.198      5.230    18.58   0.000       86.945     107.451
------------------------------------------------------------------------------
```

RQ1 - Emergence

```
. reg isEntrepreneur City_0_newyork-City_9_denver Age Created_cb_year Roll Yaw Pit

Linear regression                               Number of obs     =      6,819
                                                F(20, 6798)       =      41.42
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0973
                                                Root MSE          =     .47286
```

isEntrepreneur	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
City_0_newyork	-0.077	0.018	-4.34	0.000	-0.112	-0.042
City_1_boston	-0.096	0.026	-3.70	0.000	-0.146	-0.045
City_2_losangeles	0.055	0.030	1.83	0.068	-0.004	0.114
City_3_seattle	0.027	0.037	0.75	0.455	-0.044	0.099
City_4_washington	-0.118	0.031	-3.77	0.000	-0.179	-0.056
City_5_chicago	-0.065	0.028	-2.32	0.020	-0.120	-0.010
City_6_austin	0.058	0.041	1.44	0.151	-0.021	0.138
City_7_sandiego	0.015	0.051	0.30	0.764	-0.085	0.115
City_8_atlanta	-0.021	0.045	-0.48	0.630	-0.109	0.066
City_9_denver	-0.022	0.041	-0.54	0.589	-0.101	0.058
Age	-0.010	0.001	-15.20	0.000	-0.011	-0.008
Created_cb_year	-0.048	0.003	-18.40	0.000	-0.053	-0.043
Roll	0.002	0.002	0.81	0.421	-0.003	0.006
Yaw	0.001	0.002	0.46	0.647	-0.003	0.005
Pitch	-0.003	0.002	-1.21	0.227	-0.007	0.002
BMI	-0.002	0.002	-1.56	0.118	-0.006	0.001
Bachelor	0.024	0.015	1.66	0.098	-0.004	0.053
Master	-0.043	0.016	-2.63	0.009	-0.076	-0.011
PhD	0.071	0.025	2.83	0.005	0.022	0.120
OFA	-0.102	0.039	-2.61	0.009	-0.179	-0.025
_cons	97.198	5.230	18.58	0.000	86.945	107.451

- **Coefficient**: It **quantifies** the **strength** and **direction** of the relationship **between** an **independent** variable and the **dependent** variable.
- **Standard Error (Std Err.):** It **measures** the **average distance** that the **estimated coefficient is expected to deviate from the actual population coefficient**. A smaller standard error indicates more precise estimates.
- **t-Statistic (t):** This is a value computed to **determine if a coefficient significantly differs from zero**. A larger absolute value of the t-statistic indicates a higher likelihood that the coefficient is meaningfully different from zero in the population.
- **p-Value (p < |t|):** It's typically **testing whether the coefficient is significantly different from zero**. A small p-value (usually less than 0.05) suggests that the coefficient is statistically significant, meaning it's unlikely to be zero just by chance.
- **Confidence Intervals**: The **confidence interval for a coefficient gives a range of values within which the true population coefficient is expected to fall**, with a certain level of confidence (commonly 95%). Wider intervals indicate more uncertainty in the estimate, while narrower intervals suggest greater precision.

RQ1 - Emergence

Entrepreneurial emergence using Facial Ratios - “Entrepreneurs vs Non-Entrepreneurs” – Linear Regression.

Variable	(1)	(2)	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Age	-0.010***	-0.010***	-0.010***	-0.010***	-0.010***	-0.010***	-0.009***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Created CB Year	-0.048***	-0.048***	-0.048***	-0.048***	-0.048***	-0.048***	-0.048***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
BMI	-0.003	-0.003	-0.002	-0.002	-0.002	-0.002	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Bachelor	0.025	0.025	0.026	0.024	0.024	0.025	0.025
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Master	-0.045**	-0.045**	-0.043**	-0.043**	-0.043**	-0.042*	-0.041*
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
PhD	0.070**	0.070**	0.071**	0.071**	0.070**	0.072**	0.072**
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
fWHR		0.014				0.029	0.032
		(0.045)				(0.046)	(0.046)
Cheekbone			1.089***			1.125***	1.119***
			(0.258)			(0.260)	(0.259)
OFA				-0.102**		-0.108**	
				(0.039)		(0.039)	
CFA					-0.322**		-0.332**
					(0.104)		(0.104)
Constant	97.359***	97.352***	95.935***	97.198***	97.062***	95.703***	95.571***
	(5.238)	(5.238)	(5.248)	(5.230)	(5.231)	(5.240)	(5.242)
Observations	6,819	6,819	6,819	6,819	6,819	6,819	6,819
R-squared	0.096	0.096	0.099	0.097	0.098	0.100	0.100
Adj R^2	0.094	0.094	0.096	0.095	0.095	0.097	0.097

Robust standard errors are in parentheses

*** $p < .001$, ** $p < .01$, * $p < .05$

RQ1 - Emergence

- ***H1a:*** *A higher facial width to height ratio is positively associated with entrepreneurship emergence.*
  - ➢ fWHR is positive and insignificant => hypothesis 1a is not supported

- ***H2a:*** *A higher cheekbone prominence is positively associated with entrepreneurship emergence.*
  - ➢ Cheekbone prominence is positive and significant ($p < .01$) => hypothesis 2a is supported

- ***H3a:*** *Facial symmetry is associated with entrepreneurship emergence.*
  - ➢ Facial symmetry is significant ($p < .01$) => hypothesis 3a is supported

RQ1 - Success (Revenue)

Entrepreneurial Performance using Facial Ratios - “revenue” – Linear Regression.

Variable	(1)	(2)	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Age	3,077,213	2,918,637	3,074,338	3,076,002	3,106,163	2,889,341	2,918,563
	(2,987,324)	(2,958,394)	(3,001,006)	(3,184,822)	(3,140,795)	(3,174,375)	(3,129,176)
Created CB Year	−35,683,833*	−35,337,881*	−35,683,220*	−35,682,838*	−35,713,155*	−35,319,269*	−35,346,902*
	(16,099,625)	(15,994,754)	(16,110,957)	(16,230,485)	(16,216,723)	(16,151,798)	(16,137,453)
BMI	−7,192,562	−5,750,330	−7,215,324	−7,195,084	−7,116,248	−5,867,167	−5,806,853
	(7,237,374)	(7,581,317)	(7,307,548)	(6,863,295)	(6,836,440)	(7,367,017)	(7,344,879)
Bachelor	−27,888,078	−26,615,979	−28,012,796	−27,880,320	−28,100,126	−27,078,562	−27,261,327
	(31,604,689)	(31,317,203)	(30,701,332)	(31,778,706)	(31,642,451)	(30,756,930)	(30,652,488)
Master	2.182e + 08*	2.187e + 08*	2.181e + 08	2.182e + 08*	2.182e + 08*	2.185e + 08	2.185e + 08
	(1.111e + 08)	(1.114e + 08)	(1.120e + 08)	(1.110e + 08)	(1.110e + 08)	(1.121e + 08)	(1.121e + 08)
PhD	–32,018,957	–33,464,121	–32,061,080	–32,011,387	–32,216,628	–33,583,756	–33,752,474
	(59,956,097)	(60,460,936)	(59,807,927)	(60,573,502)	(60,366,056)	(60,858,906)	(60,660,155)
fWHR		−1.889e + 08				−1.912e + 08	−1.901e + 08
		(2.146e + 08)				(2.078e + 08)	(2.091e + 08)
Cheekbone			−27,005,287			−1.233e + 08	−1.198e + 08
			(8.570e + 08)			(7.939e + 08)	(7.893e + 08)
OFA				694,805		8,216,663	
				(1.406e + 08)		(1.413e + 08)	
CFA					−51,794,808		−24,360,723
					(3.476e + 08)		(3.520e + 08)
Constant	7.194e + 10*	7.157e + 10*	7.197e + 10*	7.194e + 10*	7.200e + 10*	7.167e + 10*	7.172e + 10*
	(3.251e + 10)	(3.240e + 10)	(3.225e + 10)	(3.278e + 10)	(3.275e + 10)	(3.243e + 10)	(3.242e + 10)
Observations	1,569	1,569	1,569	1,569	1,569	1,569	1,569
R-squared	0.081	0.082	0.081	0.081	0.081	0.082	0.082
Adj R^2	0.042	0.042	0.041	0.041	0.041	0.041	0.041

Robust standard errors are in parentheses.
*** p < .001, ** p < .01, * p < .05.

RQ1 - Success (Received Funding)

Entrepreneurial performance using facial ratios - “has the company received funding” – linear regression.

Variable	(1)	(2)	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Age	0.004***	0.004***	0.004***	0.004***	0.004***	0.004***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Created CB Year	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
BMI	. 000	-0.001	. 000	-0.001	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Bachelor	-0.039	-0.039	-0.038	-0.039	-0.039	-0.038	-0.038
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Master	-0.038	-0.038	-0.037	-0.038	-0.038	-0.038	-0.038
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
PhD	0.006	0.006	0.007	0.006	0.006	0.007	0.007
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
fWHR		0.080				0.087	0.087
		(0.062)				(0.063)	(0.063)
Cheekbone			0.412			0.458	0.460
			(0.374)			(0.376)	(0.376)
OFA				0.016		0.011	
				(0.056)		(0.056)	
CFA					0.041		0.026
					(0.150)		(0.151)
Constant	10.346	10.487	9.894	10.327	10.333	9.985	9.987
	(7.576)	(7.574)	(7.573)	(7.577)	(7.577)	(7.570)	(7.570)
Observations	2,144	2,144	2,144	2,144	2,144	2,144	2,144
R-squared	0.069	0.069	0.069	0.069	0.069	0.070	0.070
Adj R^2	0.040	0.040	0.040	0.039	0.039	0.040	0.040

Robust standard errors are in parentheses

*** $p < .001$, ** $p < .01$, * $p < .05$

RQ1 - Success (Total Funding Received)

Entrepreneurial performance using facial ratios - “total amount of funding received” - Tobit regression.

Variable	(1)	(2)	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Age	1,499,281***	1,506,924***	1,507,953***	1,498,037***	1,507,013***	1,517,512***	1,526,675***
	(330,244)	(329,946)	(330,779)	(326,257)	(327,585)	(326,734)	(328,015)
Created CB Year	−3,326,985*	−3,334,328*	−3,320,190*	−3,326,958*	−3,325,805*	−3,328,120*	−3,326,798*
	(1,467,227)	(1,469,445)	(1,469,496)	(1,467,088)	(1,467,718)	(1,471,529)	(1,472,103)
BMI	−1,447,282	−1,504,349	−1,379,129	−1,448,779	−1,434,647	−1,440,277	−1,426,363
	(871,152)	(883,825)	(876,249)	(878,306)	(878,333)	(893,502)	(892,386)
Bachelor	−11,470,627	−11,453,728	−11,368,529	−11,469,565	−11,485,951	−11,345,246	−11,359,983
	(7,287,668)	(7,291,622)	(7,306,356)	(7,287,726)	(7,289,568)	(7,311,887)	(7,312,905)
Master	−7,024,168	−7,107,686	−6,914,282	−7,024,629	−7,033,531	−7,003,894	−7,014,783
	(8,301,393)	(8,290,418)	(8,311,465)	(8,301,797)	(8,300,744)	(8,301,966)	(8,301,093)
PhD	4,484,635	4,505,046	4,692,391	4,495,203	4,406,122	4,723,856	4,638,281
	(10,133,460)	(10,142,872)	(10,186,624)	(10,089,426)	(10,099,429)	(10,153,173)	(10,161,500)
fWHR		6,834,942				7,848,516	8,058,552
		(19,864,710)				(20,104,351)	(20,127,550)
Cheekbone			64,881,914			68,760,701	70,044,016
			(1.122e + 08)			(1.139e + 08)	(1.135e + 08)
OFA				550,232		−144,305	
				(16,686,657)		(16,786,880)	
CFA					−10,612,764		−12,446,065
					(43,221,232)		(43,454,020)
Constant	6.572e + 09*	6.575e + 09*	6.484e + 09*	6.572e + 09*	6.570e + 09*	6.483e + 09*	6.478e + 09*
	(2.958e + 09)	(2.959e + 09)	(2.972e + 09)	(2.958e + 09)	(2.959e + 09)	(2.972e + 09)	(2.973e + 09)
sigma:Constant	79,267,558***	79,282,982***	79,272,237***	79,266,580***	79,269,065***	79,290,505***	79,292,503***
	(9,526,295)	(9,527,549)	(9,529,437)	(9,523,537)	(9,527,513)	(9,528,579)	(9,532,653)
Observations	2,144	2,144	2,144	2,144	2,144	2,144	2,144
pseudo-R2	0.013	0.013	0.013	0.013	0.013	0.013	0.013
LL	−6,817.845	−6,817.793	−6,817.701	−6,817.844	−6,817.822	−6,817.633	−6,817.602

Robust standard errors are in parentheses.
*** $p < .001$, ** $p < .01$, * $p < .05$.

RQ1 - Success (Valuation)

Entrepreneurial Performance using Facial Ratios - “valuation” – Linear Regression.

Variable	(1)	(2)	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Age	−14,929,903	−15,284,383	−14,814,758	−13,897,448	−14,223,215	−14,233,312	−14,580,799
	(18,890,179)	(18,723,655)	(19,431,112)	(18,186,901)	(18,527,121)	(18,467,170)	(18,804,726)
Created CB Year	−1.356e + 08*	−1.336e + 08*	−1.355e + 08*	−1.380e + 08*	−1.396e + 08*	−1.359e + 08*	−1.376e + 08*
	(62,071,790)	(63,172,623)	(62,514,728)	(64,176,755)	(64,722,557)	(65,531,531)	(66,102,530)
BMI	11,663,321	16,374,577	11,966,708	14,940,342	16,046,223	19,823,117	20,827,566
	(29,606,717)	(26,923,730)	(29,006,972)	(32,164,937)	(31,853,086)	(29,310,044)	(29,061,092)
Bachelor	99,521,711	84,884,278	1.005e + 08	1.036e + 08	90,317,122	88,790,328	75,441,496
	(2.189e + 08)	(2.107e + 08)	(2.212e + 08)	(2.183e + 08)	(2.195e + 08)	(2.117e + 08)	(2.130e + 08)
Master	2.071e + 08	2.104e + 08	2.077e + 08	2.311e + 08	2.281e + 08	2.349e + 08	2.315e + 08
	(2.553e + 08)	(2.566e + 08)	(2.569e + 08)	(2.524e + 08)	(2.567e + 08)	(2.547e + 08)	(2.591e + 08)
PhD	7.549e + 08	7.413e + 08	7.555e + 08	7.492e + 08	7.465e + 08	7.352e + 08	7.326e + 08
	(5.983e + 08)	(6.132e + 08)	(5.973e + 08)	(5.906e + 08)	(5.890e + 08)	(6.041e + 08)	(6.025e + 08)
fWHR		−6.393e + 08				−6.520e + 08	−6.467e + 08
		(8.657e + 08)				(8.476e + 08)	(8.500e + 08)
Cheekbone			4.020e + 08			25,054,722	−8,826,420
			(4.185e + 09)			(4.055e + 09)	(4.088e + 09)
OFA				−6.382e + 08		−6.496e + 08	
				(9.464e + 08)		(9.492e + 08)	
CFA					−2.322e + 09		−2.333e + 09
					(2.543e + 09)		(2.558e + 09)
Constant	2.739e + 11*	2.709e + 11*	2.732e + 11*	2.786e + 11*	2.819e + 11*	2.756e + 11*	2.789e + 11*
	(1.252e + 11)	(1.269e + 11)	(1.279e + 11)	(1.294e + 11)	(1.305e + 11)	(1.333e + 11)	(1.345e + 11)
Observations	329	329	329	329	329	329	329
R-squared	0.242	0.243	0.242	0.244	0.245	0.245	0.247
Adj R^2	0.054	0.053	0.051	0.053	0.055	0.048	0.050

Robust standard errors are in parentheses.
*** p < .001, ** p < .01, * p < .05.

RQ1

- ***H1b:*** *A higher facial width to height ratio is positively associated with firm performance.*
  - ➢ fWHR is insignificant => hypothesis 1b is not supported

- ***H2b:*** *A higher cheekbone prominence is positively associated with firm performance.*
  - ➢ Cheekbone prominence is insignificant => hypothesis 2b is not supported

- ***H3b:*** *Facial symmetry is associated with firm performance.*
  - ➢ Facial symmetry is insignificant => hypothesis 3b is not supported

RQ2

- **Hypothesis 1a (Crunchbase):** *Investors view the ventures of founders who smile more favorably than founders who do not smile.*
  - **Results:** Supported. A unit increase in the smiling variable is associated with a 25.2% increase in the odds of having raised capital and leads to $3.4 million more funding.

- **Hypothesis 1b (Shark Tank):** *The duration of smiling has an inverted U-shaped relationship with funding.*
  - **Results:** Supported. The odds of receiving funding increase by a factor of 1.47 for a standard deviation increase in the smiling score. However, the inverted U-shape is not explicitly confirmed.

- **Limitations**: the "average smiling" of the founders of a company could suffer from bias
  - **Entrepreneurs** that took their **photos AFTER receiving funding** would be more likely to smile.

Randomized Experiments, Mediation, and Moderation

Purpose of Randomized Experiments

- Establish causality between variables by controlling for confounding factors.

- Reduce bias by randomly assigning subjects to treatment or control groups.

- Enhance the reliability of results through controlled conditions.

- Provide robust data for testing hypotheses related to variables of interest.

Conducting Randomized Experiments

Designing Randomized Experiments

- **Sample Selection**: Choose a representative sample of participants.

- **Random Assignment**: Randomly assign participants to treatment and control groups to minimize bias.

- **Intervention**: Implement the experimental treatment (e.g., a smiling founder vs. a neutral expression).

- **Data Collection**: Measure outcomes (e.g., funding amounts, investor perceptions) after the intervention.

RQ2 – Randomized Experiment

- We designed a within-subjects randomized experiment to verify our previous results:
  - **10 ventures/pitch-decks** presented **to 25 venture capitalists**
  - **Pitch decks:** from active companies with two versions of each pitch deck e.g. happy and neutral facial expressions of team members
  - **VC**: from United Kingdom with $406 million of capital under management

RQ2 –Experiment Variables

- **Dependent Variables**
  - **Probability of Successful Exit:** a liquidity event for investors that involves the venture being acquired, or going public
  - **Valuation:** the investor's valuation of the company
  - **Favorability:** the extent to which the company represented an overall favorable investment opportunity

- **Independent Variable**
  - **Smiling Pitch Deck:** whether the pitch deck contained faces with smiling entrepreneurs (coded as "1" for smiling and "0" otherwise)

- Run **fixed-effects regressions** with each pitch deck as the grouping (within) variable.

RQ2 – Experiment Results

- **Hypothesis 2**. *Venture Capitalists will evaluate pitch decks depicting smiling founders more positively than pitch decks depicting non-smiling founders.*
  - Companies with **smiling pitch decks** were associated with a **higher probability of successful exit** ($p < .01$).
  - Companies with **smiling pitch decks** were associated with **higher company valuation** ($p < .01$).
  - Companies with **smiling pitch decks** were regarded as **a higher investment opportunity** ($p < .01$).

Mediation Mechanism

- **Definition:** Mediation occurs when the effect of an independent variable on a dependent variable is transmitted through a mediator variable.

- **Example:** The effect of a founder's facial expression (IV) on funding outcomes (DV) may be mediated by perceived trustworthiness (mediator).

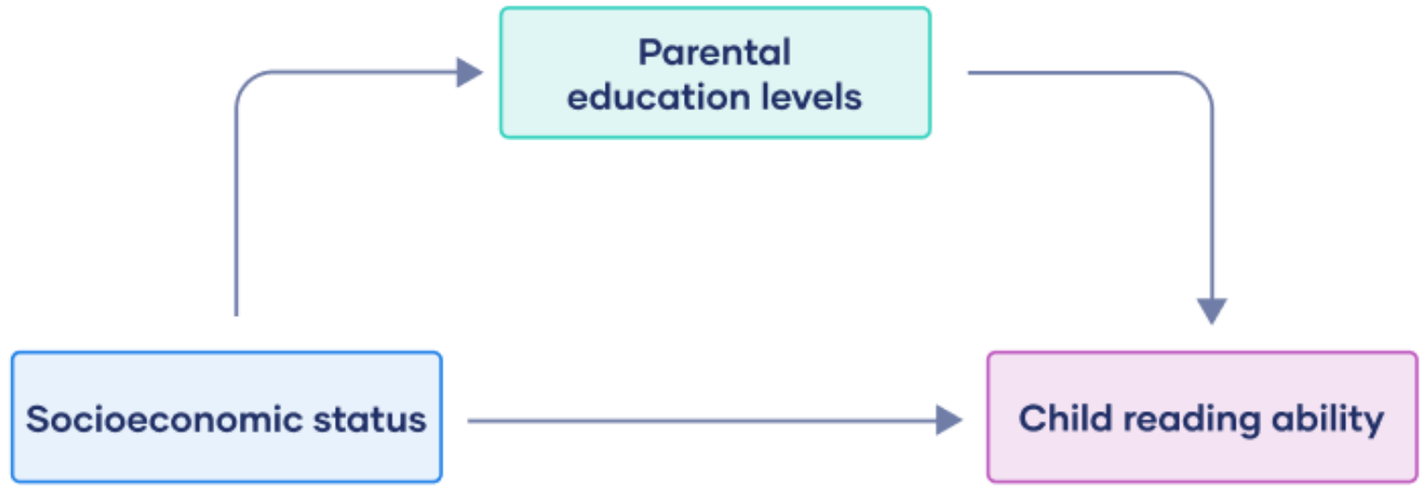
- **Significance:** Helps to clarify the process through which the independent variable influences the dependent variable.

Mediation Mechanism

- In a study on socioeconomic status and reading ability in children, you hypothesize that parental education level is a mediator.
- This means that **socioeconomic status (X) affects reading ability (Y) mainly through** its influence on **parental education levels (M)**.

- **Full mediation**: a mediator fully explains the relationship between the independent and dependent variable: without the mediator in the model, there is no relationship.

- **Partial mediation:** there is still a statistical relationship between the independent and dependent variable even when the mediator is taken out of a model: the mediator only partially explains the relationship.



RQ2 – Mediation Mechanism

- **Hypothesis 2a (Crunchbase):** *Trustworthiness mediates the relationship between smiling and investors' favorable views of entrepreneurs' ventures.*
  - **Results:** Supported. Trustworthiness was identified as a mediator in the relationship between smiling and investor funding.
- **Hypothesis 2b (Shark Tank):** *Trustworthiness mediates the relationship between the duration of smiling and investors' favorable views of entrepreneurs' ventures.*
  - **Results:** Supported. Trustworthiness mediates the relationship between smiling and the likelihood of receiving funding.

Moderation Mechanism

- **Definition:** Moderation occurs when the strength or direction of the relationship between an independent variable and a dependent variable changes depending on the level of a moderator variable.

- **Example:** The effect of a founder's facial expression on funding might be moderated by investor characteristics (e.g., risk tolerance).

- **Significance:** Identifies conditions under which relationships are stronger or weaker.

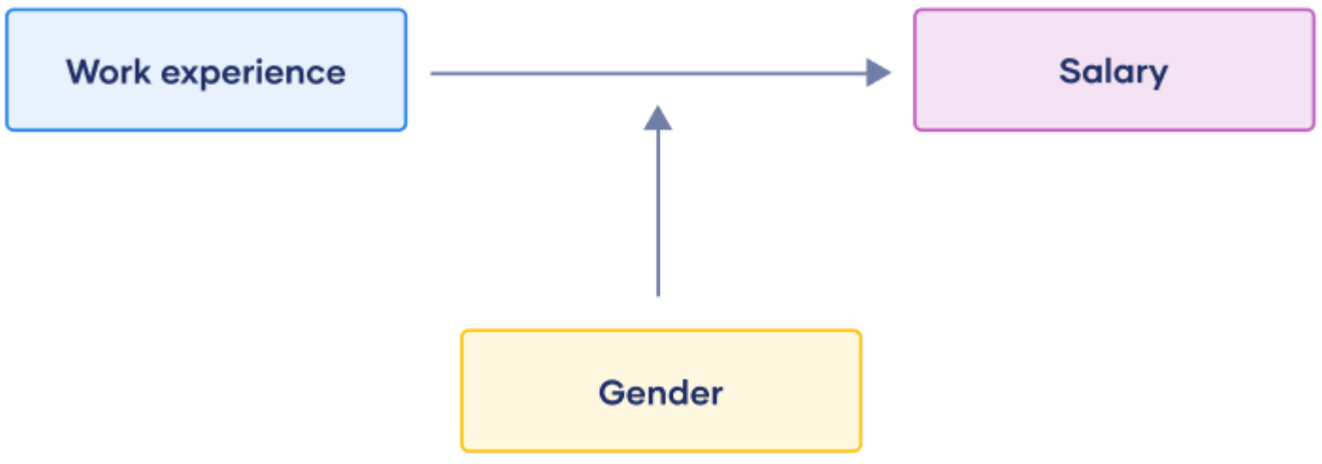
Moderation Mechanism

- In a study on work experience and salary, you hypothesize that:

❑ years of work experience predicts salary, when controlling for relevant variables,

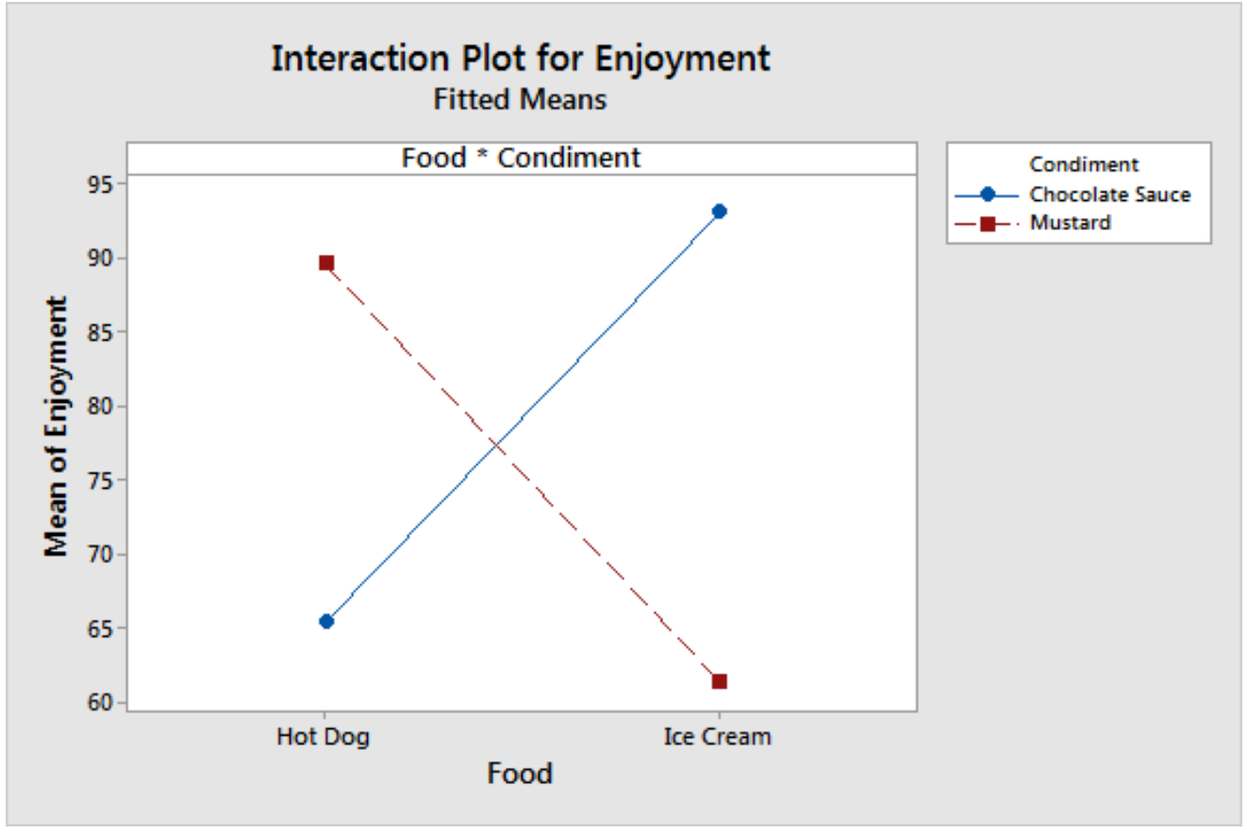
❑ gender identity moderates the relationship between work experience and salary.

- This means that the relationship between years of experience and salary would differ between men or women.



Analyzing Moderation Effects

- Examine the interaction between the independent variable and the moderator.
- Test if the relationship between the independent variable and the dependent variable changes at different levels of the moderator.



RQ2 - Gender Moderation

- **Hypothesis 3a (Crunchbase):** *Gender moderates the relationship between smiling and investors' favorable views of entrepreneurs' ventures, such that the effect of smiling on favorable views is greater for female than male entrepreneurs.*
  - **Results:** Not Supported. Gender was not found to be a significant moderator in the Crunchbase dataset.

- **Hypothesis 3b (Shark Tank):** *Gender moderates the inverted U-shaped relationship between smiling duration and investors' favorable views of entrepreneurs' ventures, such that the inflection point of the curve for female entrepreneurs, where the benefits of smiling start to decline, occurs at a higher level of smiling duration compared to male entrepreneurs.*
  - **Results:** Not Supported. Gender was not found to be a significant moderator in the Shark Tank dataset.

RQ2 - Race Moderation

- **Hypothesis 4a (Crunchbase):** *Race moderates the relationship between smiling and investors' favorable views of entrepreneurs' ventures, such that the effect of smiling on favorable views is greater for white than black entrepreneurs.*
  - **Results:** Partially Supported. Interactions for ethnicity were significant in the Crunchbase dataset.

- **Hypothesis 4b (Shark Tank):** *Race moderates the inverted U-shaped relationship between smiling duration and investors' favorable views of entrepreneurs' ventures, such that the inflection point of the curve for white entrepreneurs, where the benefits of smiling start to decline, occurs at a higher level of smiling duration compared to black entrepreneurs.*
  - **Results:** Not Supported. The interactions for ethnicity were not observed in the Shark Tank dataset.

RQ2 - Attractiveness Moderation

- **Hypothesis 5a (Crunchbase):** *Attractiveness moderates the relationship between smiling and investors' favorable views of entrepreneurs' ventures, such that the effect of smiling on favorable views is greater for more attractive entrepreneurs than for less attractive ones.*
  - **Results:** Supported. Interactions for attractiveness were significant in the Crunchbase dataset.

- **Hypothesis 5b (Shark Tank):** *Attractiveness moderates the inverted U-shaped relationship between smiling duration and investors' favorable views of entrepreneurs' ventures, such that the inflection point of the curve for more attractive entrepreneurs, where the benefits of smiling start to decline, occurs at a higher level of smiling duration compared to less attractive entrepreneurs.*
  - **Results:** Not Supported. Interactions for attractiveness were not observed in the Shark Tank dataset.

RQ2 - Additional Analysis

- **Investigating Team Composition and Smiling Variability**
  - **Team Composition and Smiling Variability (average pairwise distance):** Explored as potential moderating factors in the relationship between smiling and funding.
  - **Datasets Utilized**: Crunchbase and Shark Tank.

- **Outcomes:**
  - **Team Composition**: Investigated the dynamics between solo founders and founder teams in relation to funding.
  - **Smiling Variability**: Explored the variability of smiling within founder teams and its potential impact on funding.

RQ3 – Results

- ***RQ1:*** *Are entrepreneurs more likely than non-entrepreneurs to exhibit positive emotions on social media platforms?*
  - *Results:* The coefficient of "**isEntrepreneur**" is **positive** and **significant** ($p<0.001$) **=> entrepreneurs** are more likely to exhibit **positive emotions than non-entrepreneurs**.

- ***RQ2:*** *Are social entrepreneurs more likely than other entrepreneurs to exhibit positive emotions on social media platforms?*
  - *Results:* The coefficient of **social entrepreneurship** is **positive** and **significant** ($p<0.001$) **=> social entrepreneurs** experience a **higher emotional positivity** than **other entrepreneurs**.

- ***RQ3:*** *Are serial entrepreneurs less likely than other entrepreneurs to exhibit positive emotions on social media platforms?*
  - *Results:* The coefficient of **serial entrepreneurship** is **negative** but **NOT significant =>** we **cannot argue** that **serial entrepreneurs** are **less likely** than **other entrepreneurs** to exhibit **positive emotions**.

Robustness Tests

Purpose of Robustness Testing

Why Conduct Robustness Tests?

- To assess the reliability of the model results under different conditions.
- To ensure that the findings are not specific to a particular dataset or model specification.
- To evaluate whether results hold when varying assumptions or excluding certain data points.
- To enhance the credibility of the research conclusions.

Types of Robustness Tests

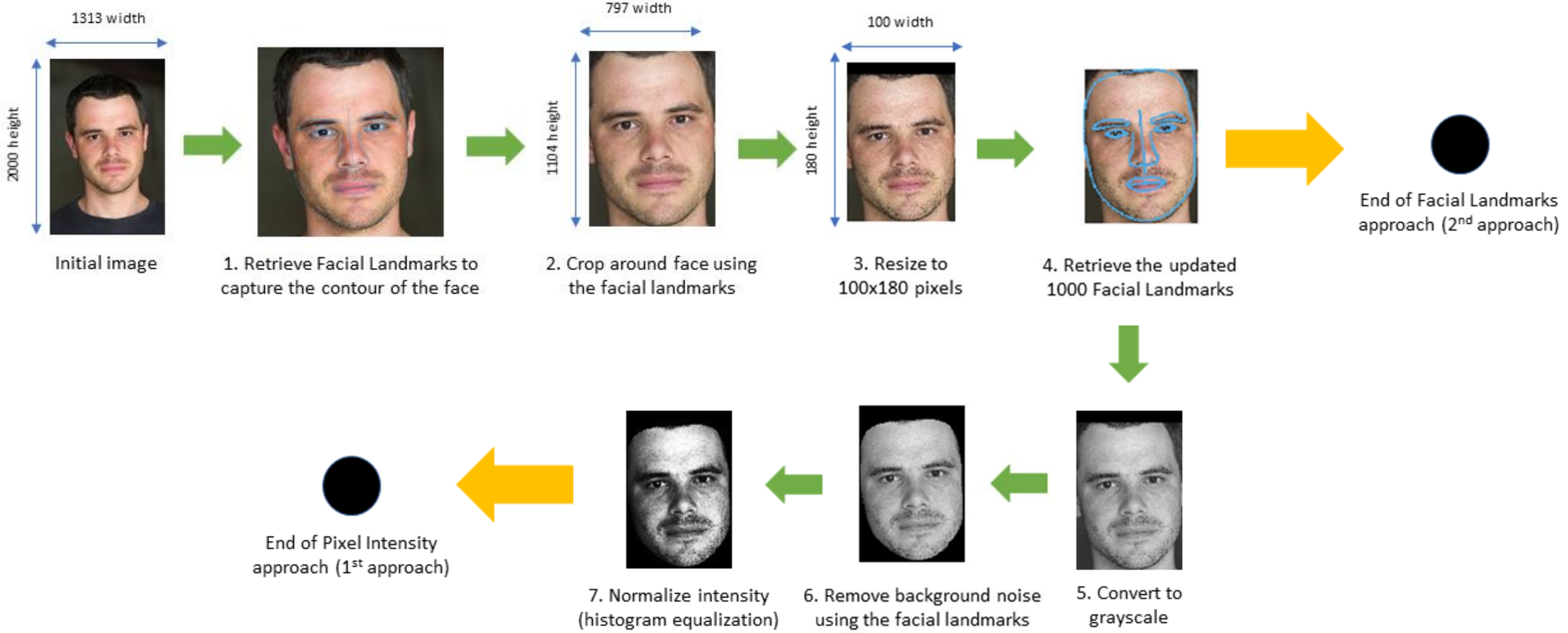
- **Sensitivity Analysis**: Examining how the output of the model changes when input parameters are varied.
- **Cross-Validation**: Splitting the data into different subsets to test the model's performance and generalizability.
- **Falsification Tests**: Testing hypotheses by evaluating whether independent variables predict outcomes in unrelated contexts (e.g., comparing entrepreneurs to managers).
- **Subgroup Analysis**: Investigating whether the model results hold for different demographic groups (e.g., by age, gender, or industry).

Robustness Test in our Research

- **Limitations of one-dimensional facial measures (e.g. fWHR):**
  - Several failures to replicate findings
  - Image variation, expressions, head pose, and face orientation may change the facial measures

- Thus, we also **examine** whether the **whole facial appearance** is related to the emergence of entrepreneurs and their firm performance.
  1. **Holistic-based approach (whole-pixels)**
  2. **Feature-based approach (facial landmarks)**

Data Preprocessing



Results

Entrepreneurship emergence using the Whole Facial Appearance - "Entrepreneurs vs Non-Entrepreneurs" - Linear Regression.

Variable	(1)	(2)
Age	-0.009***	-0.008***
	(0.001)	(0.001)
Created CB Year	-0.048***	-0.048***
	(0.003)	(0.003)
BMI	-0.002	-0.002
	(0.002)	(0.002)
Bachelor	0.032*	0.032*
	(0.015)	(0.015)
Master	-0.041*	-0.041*
	(0.017)	(0.017)
PhD	0.072**	0.071**
	(0.025)	(0.025)
Feature-based approach (facial landmarks)	0.036***	
	(0.006)	
Holistic-based approach (whole-pixels)		0.038***
		(0.006)
Constant	97.053***	97.138***
	(5.301)	(5.303)
Observations	6,639	6,639
R-squared	0.103	0.104
Adj R^2	0.101	0.101

Robust standard errors are in parentheses.
*** p < .001, ** p < .01, * p < .05.

Entrepreneurship performance using the Whole Facial Appearance - "revenue" – Linear Regression.

Variable	(1)	(2)
Age	2,802,475	3,222,712
	(2,871,535)	(3,227,454)
Created CB Year	−37,009,373*	−36,365,182*
	(16,436,531)	(16,372,838)
BMI	−6,867,784	−7,195,135
	(7,088,266)	(7,451,256)
Bachelor	−29,792,818	−26,242,721
	(32,354,488)	(31,637,306)
Master	2.138e + 08	2.168e + 08*
	(1.092e + 08)	(1.103e + 08)
PhD	−40,286,212	−34,003,345
	(62,975,354)	(60,205,522)
Feature-based approach (facial landmarks)	40,889,672	
	(29,899,720)	
Holistic-based approach (whole-pixels)		2,960,687
		(21,720,606)
Constant	7.462e + 10*	7.331e + 10*
	(3.319e + 10)	(3.306e + 10)
Observations	1,541	1,541
R-squared	0.085	0.083
Adj R^2	0.044	0.041

Robust standard errors are in parentheses.
*** p < .001, ** p < .01, * p < .05.

Results

- Both “Feature-based approach (facial landmarks)” and “Holistic-based approach (whole-pixels)” are **significant** ($p < .05$) and **positive** for emergence.

  - The **whole face is** significantly **associated** with **entrepreneurship emergence**.

- Both “Feature-based approach (facial landmarks)” and “Holistic-based approach (whole-pixels)” are **insignificant** for firm performance.

  - The **whole face is NOT associated** with **firm performance**.

Additional Robustness Tests

- **Run "falsification tests"** to investigate whether our independent variables only predict "entrepreneurial emergence" and have **no predictive power for other occupations** (e.g. managers and technical people).

1. managers vs non-managers (DV was coded as "1" for managers and "0" for non-managers)
2. technical vs non-technical people (DV was coded as "1" for technical people and "0" for non-technical people)

➢ The results show that fWHR, cheekbone prominence and OFA are not associated with other occupations.

Additional Robustness Tests

- **Compare entrepreneurs with managers**
  - There are statistically significant differences between the cheekbone prominence and facial symmetry of entrepreneurs and managers.

- **Compare the entrepreneurs with technical people**
  - There are significant differences between the cheekbone prominence and facial symmetry of entrepreneurs and technical people.

- **Compare the whole facial appearance of entrepreneurs vs managers and technical people.**
  - There are significant differences between the whole facial appearance of entrepreneurs versus managers and technical people.

RQ3 – How Job Change Affect Emotions

- **Limitation of previous results (endogeneity):**
  - We cannot establish whether **entrepreneurship leads** to the expression of **positive emotions**
    OR
  - whether **individuals** who are more likely to express **positive emotions self-select into entrepreneurship.**

- Thus, we next examined **how a job change from entrepreneur to non-entrepreneur and vice versa affects the emotions of an individual.**

1. We used the job history of entrepreneurs and non-entrepreneurs from Crunchbase and combined them with the tweets published by each user.
2. Then, we calculated the average sentiment of the tweets associated with each job.
3. Our sample consist of 37,225 job profiles and 21,491,962 tweets.

RQ3 – How Job Change Affect Emotions

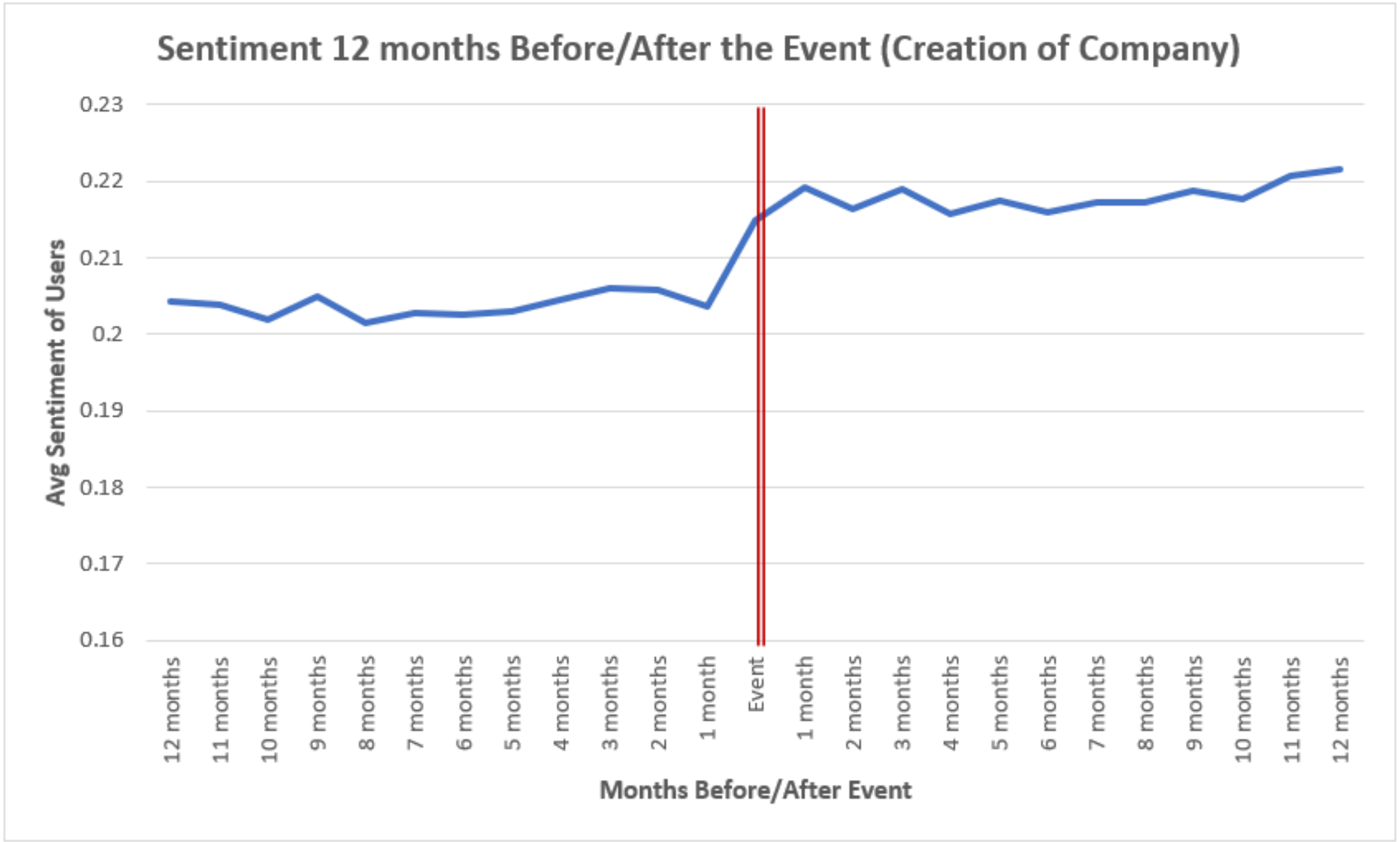


Figure: Sentiment of entrepreneurs before and after the foundation of their company.

What panel data looks like…

Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities (*i*) are observed across time (*t*).

(X_{it}, Y_{it}), $i=1,\ldots n$; $t=1,\ldots T$

These entities could be states, companies, families, individuals, countries, etc.

Entity	Year	Y	X1	X2	X3	…..
1	1	#	#	#	#	…..
1	2	#	#	#	#	…..
1	3	#	#	#	#	…..
:	:	:	:	:	:	:
2	1	#	#	#	#	…..
2	2	#	#	#	#	…..
2	3	#	#	#	#	…..
:	:	:	:	:	:	:
3	1	#	#	#	#	…..
3	2	#	#	#	#	…..
3	3	#	#	#	#	…..

The idea of fixed effects

Entity	Year	Y	X1	X2	X3	.....
1	1	#	#	#	#	.....
1	2	#	#	#	#	.....
1	3	#	#	#	#	.....
:	:	:	:	:	:	:
2	1	#	#	#	#	.....
2	2	#	#	#	#	.....
2	3	#	#	#	#	.....
:	:	:	:	:	:	:
3	1	#	#	#	#	.....
3	2	#	#	#	#	.....
3	3	#	#	#	#	.....

- **Entities** have **individual characteristics** that **may** or **may not influence the outcome** and/or predictor variables.
- **Since individual characteristics** are not random and **may impact the predictor or outcome variables**, **we need to control for them**. This way, the effect of the predictors will not be influenced by those fixed characteristics.

RQ3 – Results

- Run a **within-user (fixed-effects) analysis** in order to examine how each job (entrepreneur vs. non-entrepreneur) influenced her sentiment.

- **RQ4**: *How does a job change from entrepreneur to non-entrepreneur and vice versa affect the emotions of an individual on social media platforms?*
  - **Results:** The **coefficient** for entrepreneurship is **positive and significant** ($p<0.01$), indicating that **engagement in entrepreneurship** is associated with **more positive emotions** expressed on social media.

THANK YOU