Team 1: Deliverable 4

Aditya Agrawal : adityaai@bu.edu
Melissa Zhen : mzhen001@bu.edu
Liangzhuo Zhang : zhanglz@bu.edu
Cindy Lu : cindy19@bu.edu

Project Goal

Our goal is to collect, process, and analyze patent data of Korean and Chinese authors from patents filed in the US within the year of 2017 to 2021, so that we can then analyze gender distribution among these patent authors by what country the authors are from. We also want to see if there is a difference in gender analysis when a patent author's original character name is used in comparison to when it's translated to English.

Patents & Names Extraction Method

We used the previous team's (specifically Spring 22 Team 1) data for our analysis. The previous team's data contained information for all patents filed in the USPTO (United States Patent and Trademark Office) from 2005 to 2021. We choose patents for five years from 2017 to 2021, extracting Chinese and Korean patents from the original datasets.

The previous team had labeled the inventor's country of origin for each patent. We filtered the patents that had "CN" or "KR" in their list of inventor countries.

All patent numbers were the US Patent numbers. We use PyPI's google_patent_scraper to get the scraped Google Patents webpage for each US patent. From there we filtered worldwide applications to get the Chinese and Korean patents numbers and extract the original Chinese and Korean names of inventors from them. If a patent had both Chinese & Korean patents, we used the inventor country as a reference to select the appropriate patent.

We collected 38452 for 2017, 38768 for 2018, 50960 for 2019, 55183 for 2020 and 48597 for 2021.

Analysis of Patents & Names Extracted

After consulting with our client, we've decided to perform a gender analysis for ~8,000 Chinese names and ~8,000 Korean names for each of the years from 2017-2021. This currently gives us a sample size of ~40,000 names to work with for Chinese and Korean each.

Methods Used for Gender Prediction

English

For English names, we utilized the methods from the previous year's teams which was using Harvard Dataverse's WGND (World Gender Name Dictionary) to determine if the name was male or female. If the name was not present in the dictionary, we use Python's library "gender-guesser 0.4.0." to guess the gender, which returns {male, female, androgynous, unknown, mostly male, mostly female}.
The previous team noted, and we concur, that this might not be the best method for names that were translated from their Chinese or Korean originals, as it would often classify their genders as "Unknown."

Chinese

For Chinese names, we used an algorithm developed by Zhengxiang Wang (jaaack-wang/gender-predicator (Github)). It is a Naive Bayes algorithm that was trained on 3.65 million Chinese names to determine the gender of names. It also gives the probability of the predicted gender being correct.

Korean

For Korean names, we used an online, global name checking technology called Namsor. The Namsor API is able to extract specific information by processing and classifying names. We specifically used its "genderize" function which accepts a name as a parameter and returns the most likely gender and a calibrated probability as an accuracy score.

Gender Results Data

50 Chinese Examples

Original Name	Translated Name (English)	English Predicted Gender	Chinese Predicted Gender	Chinese Predictor Probability Value
谢云龙	Xie Yunlong	unknown	M	0.9455861571
刘秋明	Liu Qiuming	unknown	M	0.5007853122
刘秋明	Liu Qiuming	unknown	M	0.5007853122
刘秋明	Liu Qiuming	unknown	M	0.5007853122
顾嘉唯	Gu Jiawei	unknown	M	0.4964143843
崔宽峻	Choi Kwan Jun	unknown	M	0.9996198183
栗觅	Li Mi	andy	F	0.8219178082
吕胜富	Lv Shengfu	unknown	M	0.9931708452

王刚	Wang Gang	andy	M	0.9990502035
钟宁	Zhong Ning	andy	M	0.5745945946
刘俊义	Liu Junyi	unknown	M	0.9795004307
彭吉润	Peng Jirun	unknown	M	0.9281818213
马治中	Ma Zhizhong	unknown	M	0.9949888476
孙考祥	Sun Kaoxiang	unknown	M	0.9640549091
梁荣才	Liang Rongcai	unknown	M	0.984257889
王麒麟	Wang Qilin	unknown	M	0.9841868592
王文艳	Wang Wenyan	unknown	F	0.8922401953
刘万卉	Liu Wanhui	unknown	F	0.9669783151
李又欣	Li Youxin	unknown	F	0.7643847302
徐荣祥	Xu Rongxiang	unknown	M	0.9717154987
张军	Zhang Jun	mostly_male	M	0.9476945245
王晋	Wang Jin	female	M	0.8518057285
杨春燕	Yang Chunyan	unknown	F	0.9360856019
顾颖	Gu Ying	andy	F	0.9282084691
李少伟	Li Shaowei	unknown	M	0.9058931719
夏宁邵	Xia Ningshao	unknown	M	0.8421034507
龙敏	Long Min	andy	F	0.6382878399
侯冠成	Hou Guancheng	unknown	M	0.9810224421
高翔	Gao Xiang	andy	M	0.9751581369
骆仲泱	Luo Zhongyang	unknown	M	0.9510398859
岑可法	Cen Kefa	unknown	M	0.9874982336
倪明江	Ni MingJiang	andy	M	0.9686355619
宋浩	Song Hao	andy	M	0.980127318
吴卫红	Wu Weihong	unknown	M	0.5318772601
余鸿敏	Yu Hongmin	unknown	M	0.5887934158
施正伦	Shi Zhenglun	unknown	M	0.98877026
周劲松	Zhou Jinsong	unknown	M	0.9971440507
方梦祥	Fang Mengxiang	unknown	M	0.6000870611
余春江	Yu Chunjiang	unknown	M	0.8618387908
王树荣	Wang Shurong	unknown	M	0.9239950516
程乐鸣	Cheng Lemin	unknown	M	0.9081438982
王勤辉	Wang Qinhui	unknown	M	0.8480309378
H·德林	Doering Helke	mostly_female	Undefined	0.8429015144
马国刚	Wu Mingting	unknown	M	0.9998568589

F·温特	Winter Florina	female	F	0.5134231826
宋春芳	Song Chunfang	unknown	F	0.9630672177
方冬	Fang Dong	andy	F	0.5868297539
汪然敏	Wang RanMin	unknown	F	0.6677639517
夏晶俊	Xia Jing Jun	andy	F	0.7901698138
刘海龙	Liu Hailong	unknown	M	0.9844342751

(https://docs.google.com/spreadsheets/d/1U6Zs9rCdZLqeg9q8gppHGiVEXfxqtlk4QTlbDjPgl3E/edit?usp=sharing)

50 Korean Examples

Original Name	Translated Name (English)	English Predicted Gender	Korean Predicted Gender	Korean Predictor Probability Value
김동현	Kim Dong-Hyun	male	male	0.9974574497
이영훈	Lee Cheon Sook	unknown	male	0.996459741
정규열	Jung Ku Youl	unknown	male	0.9941012284
최정선	Choi Jung Sun	female	male	0.7026785306
김경남	Kim Kyung Nam	unknown	male	0.9460518074
정하진	Jung Ha Jin	unknown	male	0.7839747301
최경호	Choi Kyung Ho	male	male	0.9931845906
최영진	Choi Yeong Jin	unknown	male	0.9907905042
김일태	Kim Il Tae	unknown	male	0.9915767761
장재휘	Jang Jae Hwi	unknown	male	0.9456169617
박영재	Park Young Jae	male	male	0.9905898749
김신애	Kim Sin Ae	unknown	female	0.9419896317
김원국	Kim Won Kuk	unknown	male	0.9972063165
신효원	Sin Hyo Won	unknown	male	0.8751283279
김영식	Kim Youngsik	unknown	male	0.9974574497
심홍조	Shim Hongjo	unknown	male	0.9924272505
김윤년	Kim Yoon Nyun	unknown	male	0.6904086262
이종하	Lee Jong Ha	unknown	male	0.9946334813
박희준	Park Hee Jun	male	male	0.9940759705
박형섭	Park Hyoung Seob	unknown	male	0.9974574497
손창식	Son Chang Sik	unknown	male	0.9974574497
김태우	Kim Tae Woo	unknown	male	0.9974401301
최성일	Choi Sung Il	male	male	0.9971370384
최성일	Choi Sung Il	male	male	0.9971370384
조효성	Cho Hyo Sung	unknown	male	0.9621676536
조백환	Cho Baek Hwan	unknown	male	0.9908421254
성영경	Seong Yeong Kyeong	unknown	male	0.5030202173

김예훈	Kim Ye Hoon	unknown	male	0.9061470841
이덕운	Lee Duhgoon	unknown	male	0.9455203537
정필구	Jung Phill Gu	unknown	male	0.9921765871
정명진	Chung Myung Jin	male	male	0.9412790707
정지욱	Jeong Ji Wook	unknown	male	0.9068441642
이수열	Lee Soo Yeul	unknown	male	0.9901621179
조경일	Cho Kyung Il	male	male	0.9953892242
김동욱	Kim Dong Wook	male	male	0.9974487899
서기홍	Seo Kee Hong	unknown	male	0.9974054911
김용협	Kim Yong Hyeop	unknown	male	0.9952681586
최정선	Choi Jung Sun	female	male	0.7026785306
김경남	Kim Kyung Nam	unknown	male	0.9460518074
최경호	Choi Kyung Ho	male	male	0.9931845906
김동헌	Kim Dong Huen	unknown	male	0.9974574497
권진숙	Kwon Jin Sook	female	female	0.7154772811
안영근	Ahn Young Keun	unknown	male	0.9960389763
정명호	Jeong Myung Ho	unknown	male	0.993154043
송선정	Song Sun Jung	female	female	0.8594147426
조동련	Cho Dong Lyun	unknown	male	0.9082590716
황인홍	Hwang In Hong	male	male	0.9902254819
문경안	Moon Kyung Ahn	unknown	male	0.9430652568
공혜영	Kong Hye Young	mostly_female	female	0.96172029
권복순	Kwon Bok Soon	female	female	0.6215433656

(https://docs.google.com/spreadsheets/d/1U6Zs9rCdZLqeg9q8gppHGiVEXfxqtlk4QTlbDjPgl3E/edit?usp=sharing)

Gender Results Visualization and Analysis

Original Chinese Names

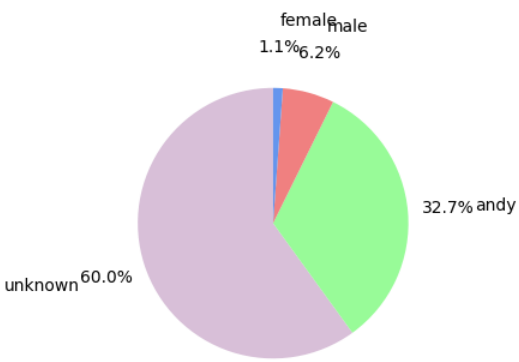
Gender Distribution for Original Chinese Names (%)				
Year	Male	Female	Androgynous	Undefined
2017	79.16 (6335)	17.43 (1395)	3.27 (262)	0.14 (11)
2018	80.09 (6411)	17.56 (1406)	2.24 (179)	0.11 (9)
2019	79.78 (6385)	17.42 (1394)	2.67 (214)	0.12 (10)
2020	79.83 (6387)	18.44 (1475)	1.6 (128)	0.14 (11)
2021	79.3 (6348)	19.21 (1538)	1.3 (104)	0.19 (15)
2017-2021	79.63 (31866)	18.01 (7208)	2.22 (887)	0.14 (56)

Translated CN Names in English

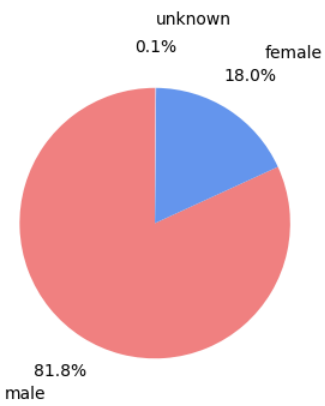
Gender Distribution for Translated Chinese Names (%)				
Year	Unknown	Androgynous	Male	Female
2017	59.98 (4800)	32.71 (2618)	6.16 (493)	1.14 (92)
2018	62.07 (4969)	30.64 (2453)	6.02 (482)	1.26 (101)
2019	61.67 (4937)	30.49 (2441)	6.19 (496)	1.61 (129)
2020	62.49 (5000)	30.92 (2474)	5.21 (417)	1.19 (110)
2021	62.2 (4977)	31.06 (2485)	5.35 (428)	1.44 (115)
2017-2021	61.68 (24683)	31.16 (12471)	5.79 (2316)	1.37 (547)

2017

English Name

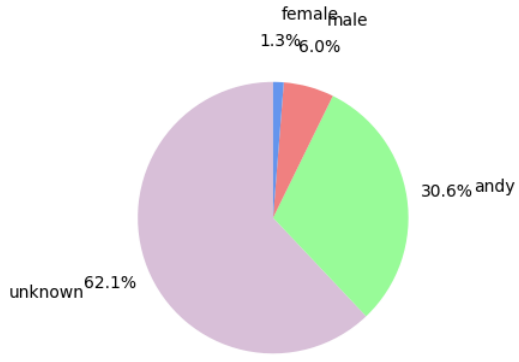


Chinese Name

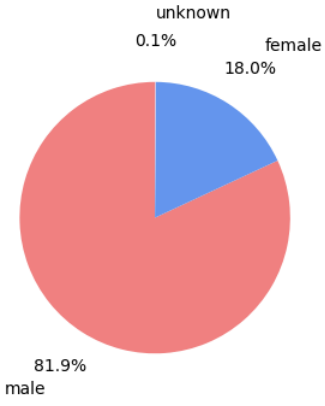


2018

English Name

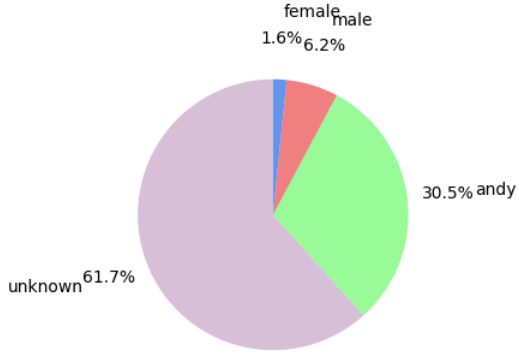


Chinese Name

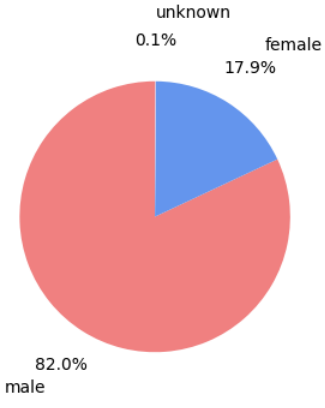


2019

English Name

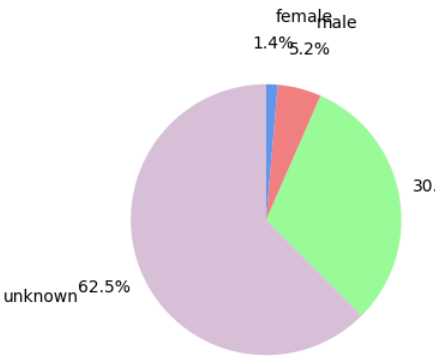


Chinese Name

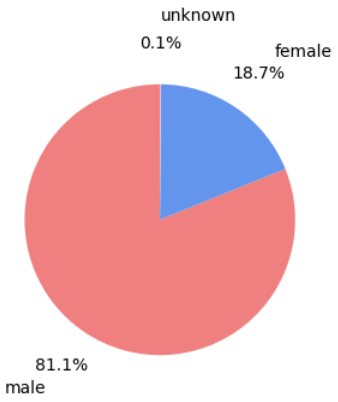


2020

English Name

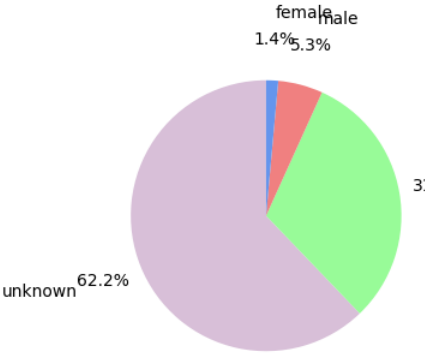


Chinese Name

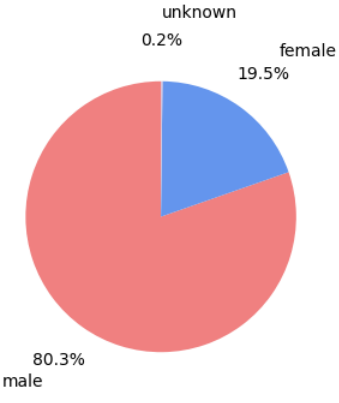


2021

English Name

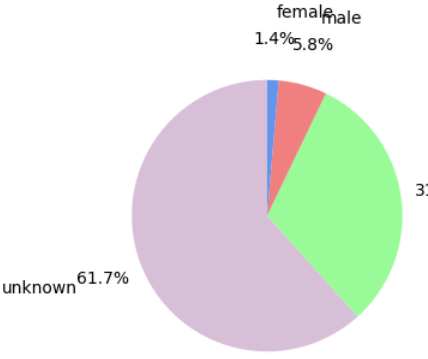


Chinese Name

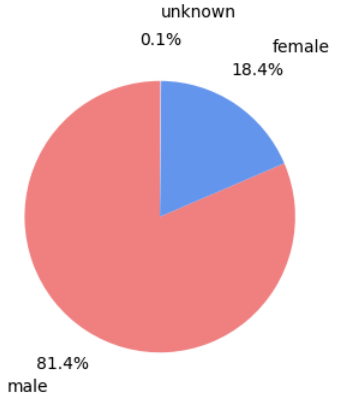


For all

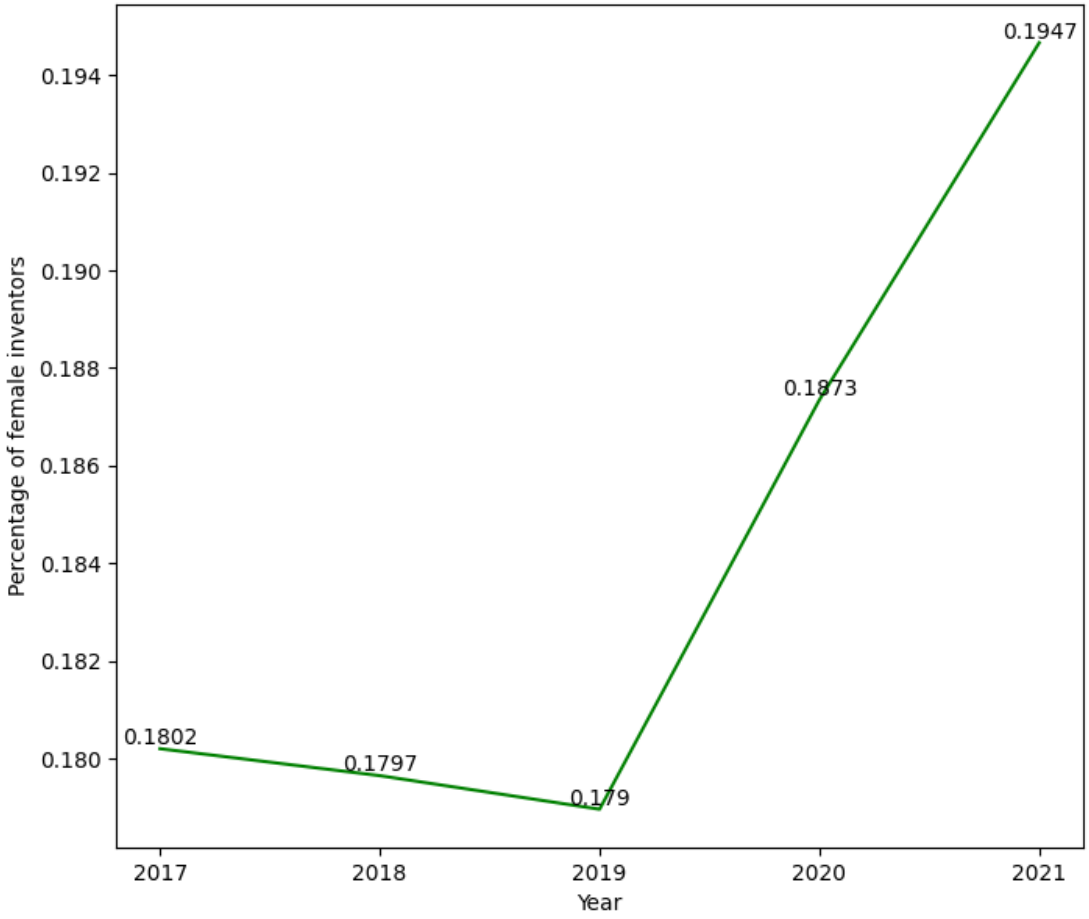
English Name



Chinese Name



The percentage of female inventors from 2017 to 2021



Original Korean Names

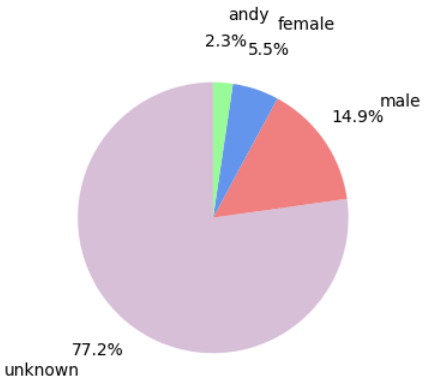
Gender Distribution for Original Korean Names (%)		
Year	Male	Female
2017	90.29 (7223)	9.71 (777)
2018	91.04 (7283)	8.96 (717)
2019	90.93 (7275)	9.07 (726)
2020	89.79 (7184)	10.21 (817)
2021	90.78 (7267)	9.22 (738)
2017-2021	90.56 (36232)	9.44 (3775)

Translated KR Names to English

Gender Distribution for Translated Korean Names (%)				
Year	Unknown	Androgynous	Male	Female
2017	75.44 (6178)	12.7 (1080)	1.5 (601)	0.35 (141)
2018	78.96 (6317)	12.83 (1026)	6.77 (543)	1.44 (115)
2019	78.83 (6307)	12.39 (991)	7.29 (583)	1.5 (120)
2020	78.29 (6264)	12.36 (989)	7.76 (621)	1.58 (127)
2021	80.23 (6419)	11.84 (947)	6.32 (506)	1.66 (133)
2017-2021	78.7 (31485)	12.58 (5033)	7.13 (2853)	1.59 (636)

2017

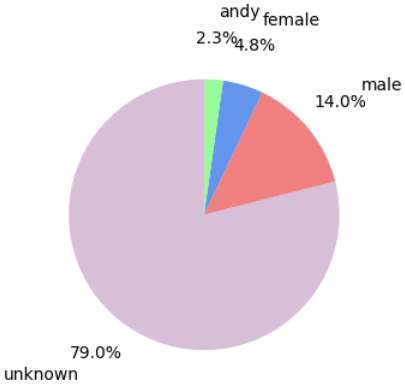
English Name



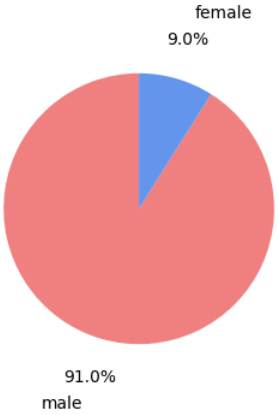
Korean Name

2018

English Name

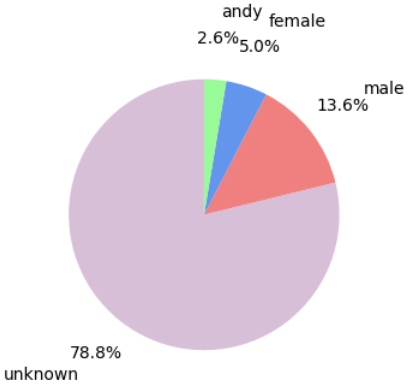


Korean Name

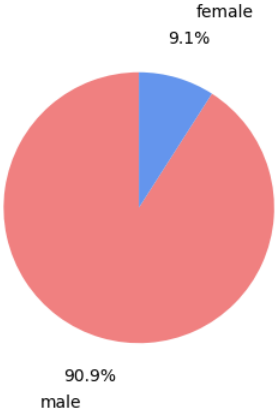


2019

English Name

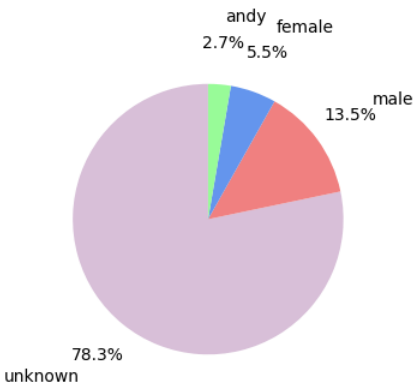


Korean Name



2020

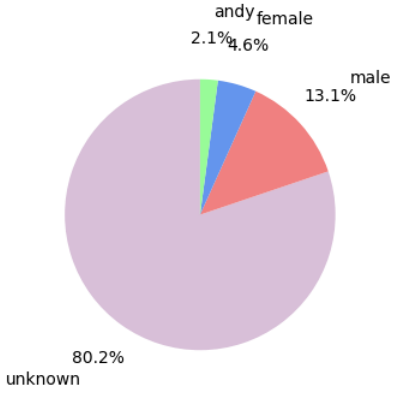
English Name



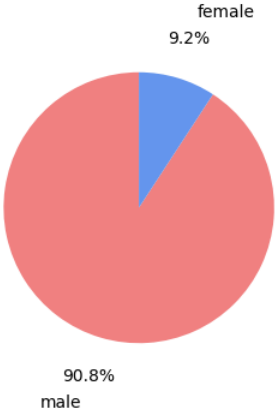
Korean Name

2021

English Name

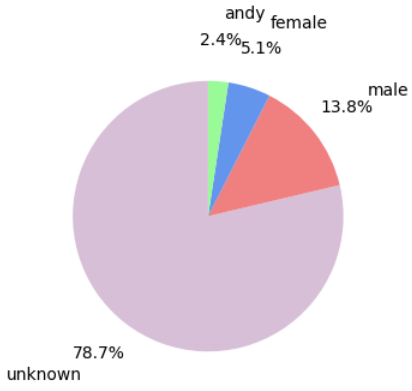


Korean Name

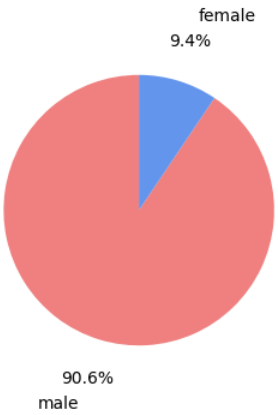


For all

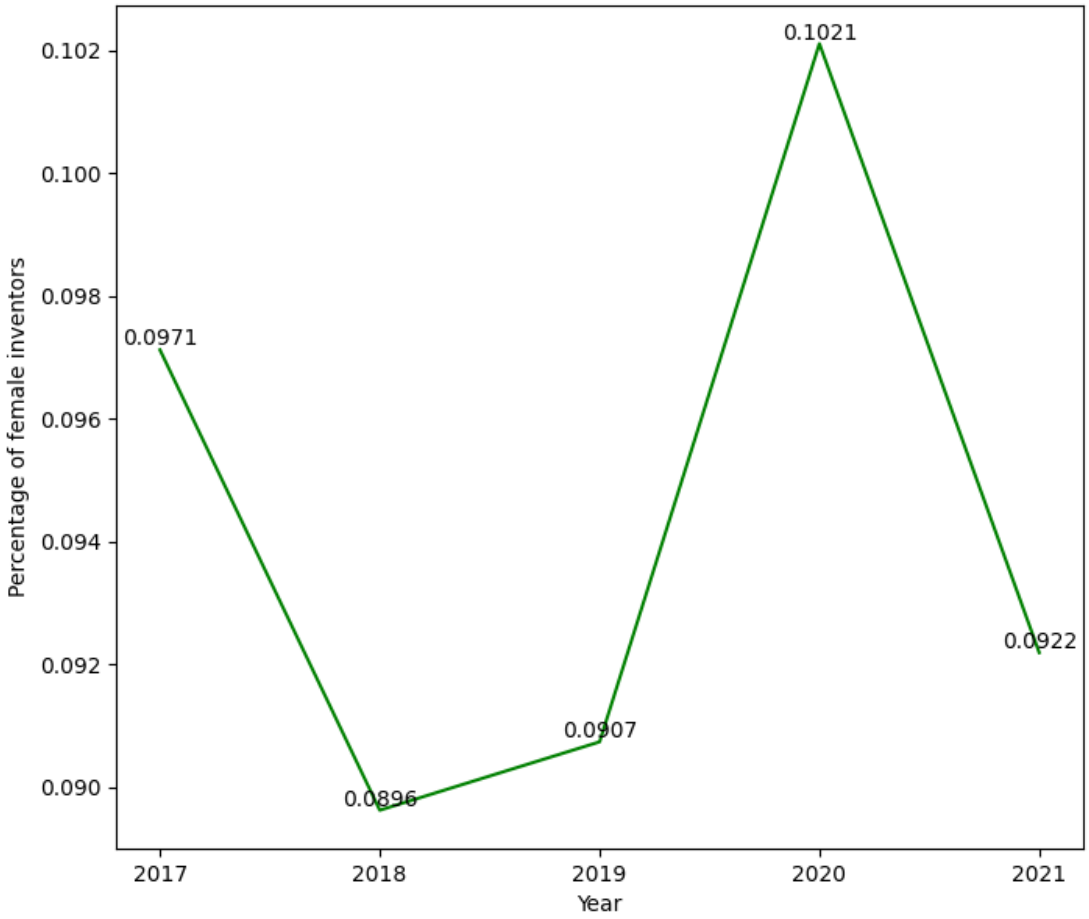
English Name



Korean Name



The percentage of female inventors from 2017 to 2021



Summary of Methods

Country	Predicted Gender	Harvard's World Gender Name Dictionary (WGND) - Translated Names	Python's gender_guesser library - Translated Names	WGND + gender_guesser - Translated Names	jaaack-wang/gender-predicator - Original Chinese Names	Namsor API - Original Korean Names
Korean (KR)	Male	0.0025 % (1)	13.80 % (5522)	13.81 % (5523)	N/A	90.56 % (36232)
	Female	0 % (0)	5.08 % (2035)	5.08 % (2035)	N/A	9.44 % (3775)
	Unknown	99.9975 % (40006)	78.7 % (31486)	78.69 % (31485)	N/A	N/A
	Androgynous	N/A	2.41 % (964)	2.41 % (964)	N/A	N/A
Chinese (CN)	Male	0 % (0)	5.78 % (2316)	5.78 % (2316)	79.63 % (31866)	N/A
	Female	0 % (0)	1.37 % (547)	1.37 % (547)	18.01 % (7208)	N/A
	Unknown	100 % (40017)	61.68 % (24683)	61.68 % (24683)	2.22 % (887)	N/A
	Androgynous	N/A	31.16 % (12471)	31.16 % (12471)	0.14 % (56)	N/A

Conclusion

In response to the first question, we found that the main difference in analyzing original characters instead of US translated names is that much of the resulting translations are not found within the gender dictionaries. Therefore most names are classified as androgynous or unknown. Gender predictors designed for English names also do not handle translated names well. Gender predictors designed for the original characters are able to predict the gender of names with very high probabilities.

In response to the second question, we found that there is an overwhelming bias towards males in the gender distribution of both Chinese inventors and Korean inventors. For Chinese inventors, about 80% of analyzed inventors were categorized as male, and this trend did not change throughout the analyzed years. For Korean inventors, we see the same trend occurring also. About 90% of the inventors were categorized as male throughout the years 2017-2021. It is worthwhile to note, however, that for both the English translated Korean and Chinese names, there are more names that are classified as unknown than male and female combined. For English translated Chinese names, there are more names classified as androgynous than male and female combined.

Previous research into gender equity of inventors in the countries of the world would often show China & Korea having a large percentage of female inventors compared to men. The reason for that is the use of translated names. But by using the original names, we see there is a higher percentage of male inventors compared to female inventors, like the rest of the world. So there still needs to be progress made in these countries to achieve gender equality for inventors.

Limitations

The data we collected from Google Patents does not contain the gender of the inventors. So we can not compare our predictions with the real gender of inventors. We can only compare the most likely accuracy of the generated genders.

Also, because we wanted to find patents in the USPTO dataset, as well as the equivalent Chinese and Korean patent offices, we utilized the patent numbers to cross-match patents across the multiple datasets. However, we soon realized that not all Chinese and Korean patents in the USPTO dataset have a corresponding Chinese or Korean version on Google Patents. Therefore, not all the data we collected is not representative of the entire Chinese and Korean patent dataset.

Additionally, because scraping and processing each year's dataset was taking an overwhelming large amount of time, we were given approval by our client to sample a subset of each year's data from the most recent 5 years–2017 to 2021. There could be unseen trends from previous years that we did not sample from.

Contributions

Aditya: Extracted patent names from 2019 and 2020. Worked on the Korean Gender Predictors and Chinese Gender predictors. Added summary of methods.
Cindy: Extracted names from 2018 patents, graphs + visualizations for Korean names, data charts for Korean and Chinese names, writing deliverables

Leon: Extracted patent names from 2021. Explored possible methods for gender prediction on names. Worked on three deliverables. Kept contact with the client and confirmed project details and problems we met with clients. Visualized the predictions to show the percentages and the trends in the proportion of women among the five years and completed relevant sections of deliverables.
Melissa: Extracted original CN & KR names from the year 2017, worked on generating the visualizations for the gender distribution of Korean and Chinese names, summarized the gender distributions for each year by number and percentages, kept contact with our client to update and ask questions whenever necessary.