Exploration Data Analysis of Korean Pop (K-pop) Idols in the Past 20 Years

Introduction to the Research Space:

Being a K-pop idol is no easy feat. There are many that train for years in order to have a spot in the limelight. Out of those thousands of people, only a heavily selected 1-2% get to debut. Debuting is when an idol trainee 'graduates' to become an active idol that is promoted by their agency company. These trainees go through years of hardcore skill training, and intense scrutiny and judgement by mentors before getting selected. The main objective of this research is to analyse trends amongst K-pop idols.

Aims and Objectives:

This research will be focusing on the general trends of K-pop idols' gender distribution, age distribution, physical traits, nationality, and their agency companies. We aim to discover the optimum qualities one may possess in order to be a successful K-pop idol, as per the South Korean entertainment industry standards.

Acquire a Dataset:

The dataset we will be referring to is the 1700+ K-pop Idols Dataset taken from Kaggle, a website containing various datasets shared by different users in its community. The link to the dataset is https://www.kaggle.com/datasets/nicolsalayoarias/all-kpop-idols/data. This will facilitate our comprehensive and more in-depth analysis of K-pop idols over the past 20 years.

Utilizing the Dataset:

We will be using Jupyter Notebook to analyse the dataset using Python, and various data visualisations will be generated so that trends can be easier seen and interpreted. These findings will be presented in a comprehensible manner such that individuals of varying technical knowledge are able to interpret them.

Writing Style to Communicate Ideas and Concepts:

This research will have concise segments that analyse and explain different parts of the dataset. This will make it easier for everyone to read the report. Additionally, technical terminologies will be minimised to ensure accessibility of comprehension amongst people across varying technical knowledge.

Summary of the Area of Research:

This research will examine different features about K-pop idols that allowed them to face the limelight. We will mainly be focusing on external features and information that had been previously shared to the public, such as height, weight, nationality, birthday, and their agency. These are some critical aspects that may come into play when a trainee is selected to debut as a K-pop idol.

Project Background

Reason for Choosing This Field of Research:

Despite not being from South Korea, I have been exposed to the Korean entertainment industry since I was young. I grew up listening to K-pop music and watching Tagalog-dubbed Korean dramas. I looked up to many artists, many of whom became my role model. There was a point in my life whereby I aspired to be a K-pop idol- I was dazzled by my favourite idols dancing and singing that I wanted to be like them too. However, reality hit and I realised that being an idol may not be as easy or achievable as I thought. I have always been curious as to what it takes for someone to be one. Although I could probably never find out the official criterias, I thought it would be interesting to analyse trends of idols' characteristics and finding out the optimum traits based on that.

Scope of Work:

We would be analysing data that will relate to an individual's physical appearance. We will not be analysing skills such as dancing and singing. We may analyse the countries where individuals come from and how that affects them being K-pop idols.

Relevancy of Dataset

Origin of data:

The dataset we will be using comes from Kaggle, a data science platform that contains community-made datasets. The author of the dataset we used goes by the name Nicolás Alayo (nicolsalayoarias). The link to this dataset is https://www.kaggle.com/datasets/nicolsalayoarias/all-kpop-idols/data.

Why This Data is Appropriate for Our Research:

This data is appropriate as it contains data of potential features that contribute to our aim- finding the optimum qualities for an individual to debut as a K-pop idol, according to the data of the idols who have debuted over the past 20 years. This includes data of height, weight, country of origin, age, and possibly their agency company. K-pop idols are heavily scrutinised on their physical appearances, which the features above heavily take a part in. Additionally, how idols are trained in order to debut relies on their agency company.

We chose this dataset over two other datasets, both of which contain data which relate to our topic. All of the datasets were from Kaggle. The first one was made by someone named Rohit Nub (onlyrohit). The link to their dataset is https://www.kaggle.com/datasets/onlyrohit/all-kpop-idols. While their dataset is very comprehensive and had a lot of idols' data, there were a few idol characteristics that it lacked compared to our dataset of choice, such as the idols' agency company and their second country. While these may seem small and does not directly give us insight towards our aim, these characteristics may give us a deeper understanding to the trends that we will see upon analysis.

The second one was made by someone named Datartist (kimjihoo). The link to their dataset is https://www.kaggle.com/datasets/kimjihoo/kpopdb. Their dataset is clearly labelled and well put together, however it lacked some essential information that we needed. These information were the height and weight of the idols. Although it is sensitive data, it plays a big role in terms of physical appearance- which is likely to influence how a trainee gets chosen to debut to be an idol.

Preparation of Dataset

Initialisation of Packages and Dataset:

In [594…

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import datetime as datetime
```

In [595…

```
#loading of dataset
#source: https://www.kaggle.com/datasets/nicolsalayoarias/all-kpop-idols/da
idolstats = pd.read_csv('kpopidolsv3.csv')
```

Initial Checks to Dataset:

We randomly generate the data of 5 idols below. We can see that there are 16 columns, excluding the index.

In [599…

```
#sample of 5 idols' data in random
idolstats.sample(5)
```

Out[599]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company
1037	Minjung	Cha Minjung	차민정	민정	15/07/2004	NaN	NaN	NaN
1292	Seungyeop	Choi Seungyeop	최승엽	승엽	8/05/1997	E'LAST	9/06/2020	E Entertainment
349	Gaga	Lee Soobin	이수빈	가가	10/11/1994	Gate9	26/01/1999	JYP| SidusHQ
451	Hayun	Choi Hayun	최하윤	하윤	16/01/2004	Hi-L	11/08/2021	Kpop Live
48	Babysoul	Lee Soojung	이수정	베이비 소울	6/07/1992	Lovelyz	17/11/2014	Woollim

From the above table, we can see that there are null values on the 'Group', 'Debut', 'Company', 'Second Country', 'Height', 'Weight', 'Birthplace', 'Other Group', and 'Former Group'. We can also see that there are idols that debuted more than 20 years ago, which is outside of our scope of research.

In [600…

```
idolstats.count()
```

Out[600]:

```
Stage Name        1778
Full Name         1769
Korean Name       1768
K Stage Name      1777
Date of Birth     1776
Group             1632
Debut             1632
Company           1632
Country           1778
Second Country      62
Height             836
Weight             566
Birthplace         834
Other Group        140
Former Group       264
Gender            1778
dtype: int64
```

From the above, we can see that out of a total of 1778 entries, almost all of the columns have null values as they do not total up to 1778. This may be due to the fact that some of these idols' information such as weight, height, and birthplace have not been publicly revealed. This may also be due to the fact that certain features do not apply to some of the idols, such as being involved in another group, being managed under an agency company, and coming from a second country. We will delve deeper into these statistics in our technical exploration later.

Ensuring the Data in Dataset is Within Our Research Scope:

Since our research will only focus on idols that have debuted in the past 20 years, let us check if our dataset matches this criteria.

In [601…

```
#converting the date layouts so that we can work with them
idolstats['Debut'] = pd.to_datetime(idolstats['Debut'], errors = 'coerce',
idolstats['Date of Birth'] = pd.to_datetime(idolstats['Date of Birth'], err
```

In [602…

```
#sorting the data in ascending order from their date of debut
idolstats.sort_values('Debut')
```

Out[602]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Country	Se Cou
1499	U-Know Yunho	Jung Yunho	정윤호	유노윤호	1986-02-06	TVXQ	1995-09-06	MBK| Turbo Co.	South Korea	
988	Max Changmin	Shim Changmin	심창민	최강창민	1988-02-18	TVXQ	1995-09-06	MBK| Turbo Co.	South Korea	
261	Dongwan	Kim Dongwan	김동완	동완	1979-11-21	Shinhwa	1998-03-24	SM| Good| Shinhwa	South Korea	
1050	Minwoo	Lee Minwoo	이민우	민우	1979-07-28	Shinhwa	1998-03-24	SM| Good| Shinhwa	South Korea	
23	Andy	Lee Sunho	이선호	앤디	1981-01-21	Shinhwa	1998-03-24	SM| Good| Shinhwa	South Korea	
...	...	...	...	...	...	...	...	...	...	
1708	Youngmin	Lim Youngmin	임영민	영민	1995-12-25	NaN	NaT	NaN	South Korea	
1714	Yubin	Cho Yubin	조유빈	유빈	1999-10-09	NaN	NaT	NaN	South Korea	
1725	Yujeong	Kim Yujeong	김유정	유정	1992-02-14	NaN	NaT	NaN	South Korea	
1740	Yulhee	Kim Yulhee	김율희	율희	1997-11-27	NaN	NaT	NaN	South Korea	
1776	Z-UK	Jeong Jaewook	정재욱	지욱	1993-01-27	NaN	NaT	NaN	South Korea	

1778 rows × 16 columns

From the above, we can see that there are anomalies in the data whereby a group is stated to have debuted in 1995, as well as null values. We will be removing these data as they are outside our research scope and will compromise our technical analysis of the data.

In [603…

```
#making a new variable for a modified dataset containing only idols who debu
df_idols = idolstats[idolstats['Debut'] >= '2004-01-01']
```

In [604…

```
#displays a table of artists who debuted over the past 20 years
df_idols.sort_values('Debut')
```

Out[604]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Country	Sc Cc
1191	Ryeowook	Kim Ryeowook	김려욱	려욱	1987-06-21	Super Junior	2005-11-06	SM	South Korea	
467	Henry	Henry Lau	헨리 라우	헨리	1989-10-11	Super Junior-M	2005-11-06	SM	Canada	-
1643	Yesung	Kim Jongwoon	김종운	예성	1984-08-24	Super Junior	2005-11-06	SM	South Korea	
453	Heechul	Kim Heechul	김희철	희철	1983-07-10	Super Junior	2005-11-06	SM	South Korea	
909	Kyuhyun	Cho Kyuhyun	조규현	규현	1988-02-03	Super Junior	2005-11-06	SM	South Korea	
...	...	...	...	...	...	...	...	...	...	
1124	Nova	Maria Pavlovna Emelyanova	마리아 파블로브나 에멜야노바	노바	2002-10-10	X:IN	2023-04-11	Escrow	Russia	
33	Aria	Baby Gauthami	베이비 고타미	아리아	2003-03-12	X:IN	2023-04-11	Escrow	India	
1172	Roa	NaN	NaN	로아	NaT	X:IN	2023-04-11	Escrow	South Korea	
278	E.sha	Kwon Yena	권예나	이샤	1998-12-29	X:IN	2023-04-11	Escrow	Australia	
149	Chi.u	Lee Jaeyi	이재이	치유	1998-08-16	X:IN	2023-04-11	Escrow	South Korea	

1604 rows × 16 columns

The table above shows the modified dataset containing only idols that have debuted in the past 20 years. We made a new variable 'df_idols' to contain them.

In [605…

```
#verifying if we have removed null values under 'Debut'
idol_stats_count = df_idols.count()
print(idol_stats_count)
```

```
Stage Name        1604
Full Name         1596
Korean Name       1595
K Stage Name      1603
Date of Birth     1602
Group             1604
Debut             1604
Company           1604
Country           1604
Second Country      56
Height             758
Weight             517
Birthplace         733
Other Group        134
Former Group       133
Gender            1604
dtype: int64
```

From the table above, the number of data in the 'Debut' column matches the number of data in the 'Stage Name' and 'Gender' columns. We can thus verify that we have successfully filtered out null values in the 'Debut' column and data that is out of our research scope. To further clean up our dataset, we will be removing any unnecessary data that will not contribute towards our aim, such as the idols' birthplace, other group, and former group.

In [606…

```
#removing any unnecessary data that do not contribute to our aim
df_idols = df_idols.drop('Birthplace', axis = 1)
df_idols = df_idols.drop('Former Group', axis = 1)
df_idols = df_idols.drop('Other Group', axis = 1)
```

In [608…

```
#verifying if the above has been successfully done
df_idols.sample(5)
```

Out[608]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Coun
52	Baekah	Kim Sua	김수아	백아	1999-10-24	XUM	2020-09-22	A100	So Ko
20	Amber	Amber Josephine Liu	엠버 조세핀 리우	엠버	1992-09-18	f(x)	2009-09-05	SM	U
1048	Mint	Kunpat Phonpawiworakun	쿤팟 퍼언빠위워라쿤	민트	1994-06-23	Tiny-G	2012-08-23	GNG	Thaila
194	Dajeong	Jung Dajeong	정다정	다정	2003-07-31	Pixy	2021-02-24	ALLART	So Ko
1699	Youngbin	Jeong Youngbin	정영빈	영빈	1998-10-05	Luminous	2021-09-09	Barunson	So Ko

From the table above, we can see that the columns for 'Birthplace', 'Other Group', and 'Former Group' have been deleted. Our dataset now contains only data that we will use to do our analysis that contribute to our aim. This makes our dataset more robust and reliable. Thus, we are now ready to do a technical analysis of our dataset.

Ethics of Use of the Dataset:

The dataset above had a licensing of CC1.0: Public Domain. There is no copyright and we are legally able to copy, modify, and distribute data without permission from the original creator.

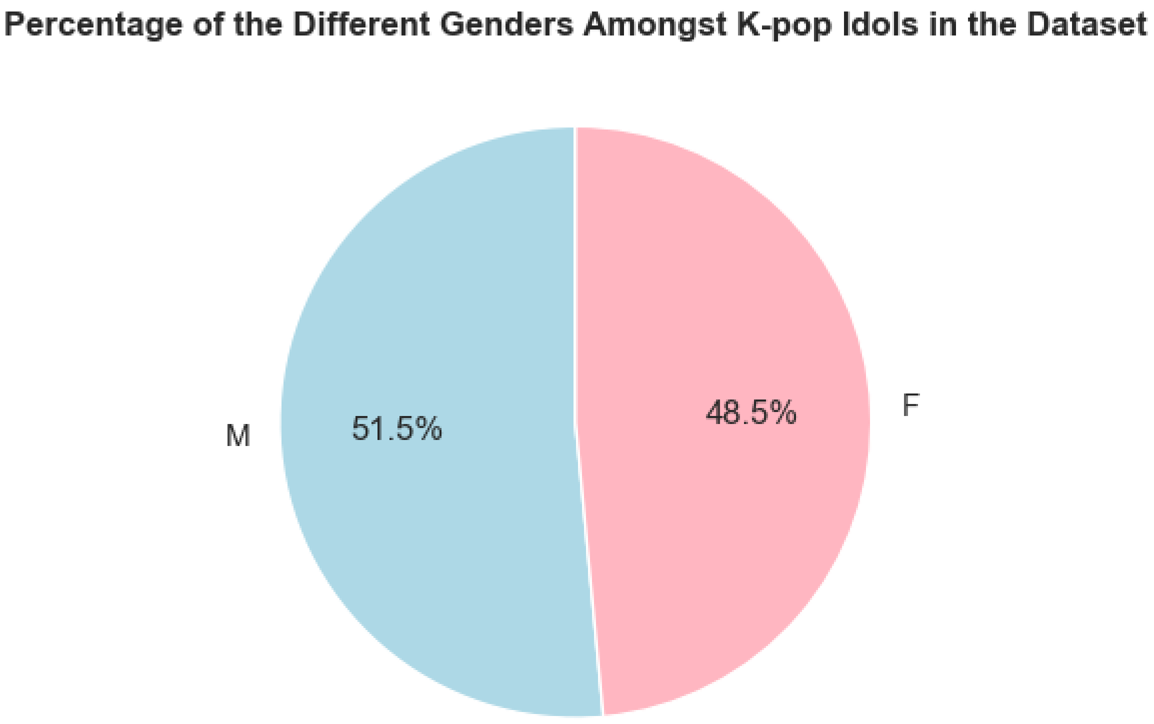
Technical Analysis of the Dataset

Gender Distribution:

First and foremost, we will analyse the gender distribution of the idols that have debuted.

In [609…

```
#plotting a pie chart to illustrate the gender distribution of K-pop idols i
labels = ['M', 'F']
colors =['lightblue', 'lightpink']

#getting the number of idols in the respective genders
diff_genders = df_idols['Gender'].value_counts()

#plotting of pie chart
plt.pie(diff_genders, labels = labels, colors = colors, startangle = 90, au

#labeling of pie chart
plt.axis('off')
plt.title('Percentage of the Different Genders Amongst K-pop Idols in the Da

plt.show()
```



From the pie chart above, we see that there are more male idols at 51.5% than female idols which was at 48.5%. This tells us that there is not a high chance of the South

Korean idol entertainment industry having gender biases in terms of debuting trainees as the statistics above are almost equal.

Idol Country Distribution:

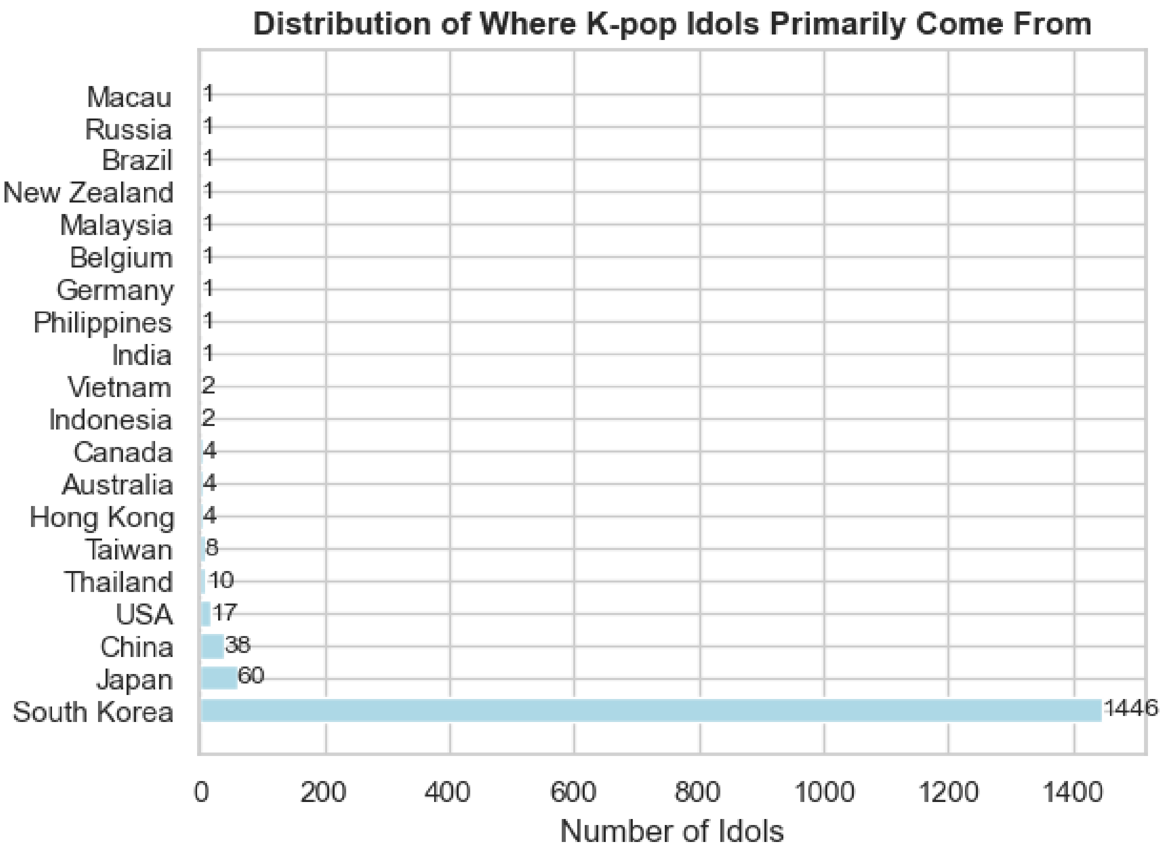
Now that we have determined the gender distribution of the idols in our dataset, let us analyse more closely on where these idols are primarily from.

In [610…

```
#plotting a horizontal bar graph depicting the different countries idols car
#and the distribution of idols in them

#getting the number of idols from the respective countries
idol_countries = df_idols['Country'].value_counts()

#plotting the horizontal bar graph
plt.barh(idol_countries.index, idol_countries, color = 'lightblue')

#adding the labels depicting the exact count of idols from their respective
for index, value in enumerate(idol_countries):
    plt.text(value, index, str(value), va = 'center', ha = 'left', fontsize

#label and title of the graph
plt.xlabel('Number of Idols')
plt.title('Distribution of Where K-pop Idols Primarily Come From', fontweigl
```

Out[610]:

```
Text(0.5, 1.0, 'Distribution of Where K-pop Idols Primarily Come From')
```



From the horizontal bar chart above, we can clearly see that the vast majority of K-pop idols come from South Korea, with 1446 idols. This is as expected as it is logical that South Korean individuals would prefer to stay in their country given that K-pop originated there in the first place. Additionally, since most K-pop idols sing and mainly converse in Korean, there will also be less language barrier for South Koreans. It is interesting to note that the countries that came in second and third were Japan and China respectively. Geographically, these countries are the nearest to South Korea, and they also have languages that look and sound similar to Korean due to their common historical and cultural interactions. Hence, the trends shown on the above data is as expected.

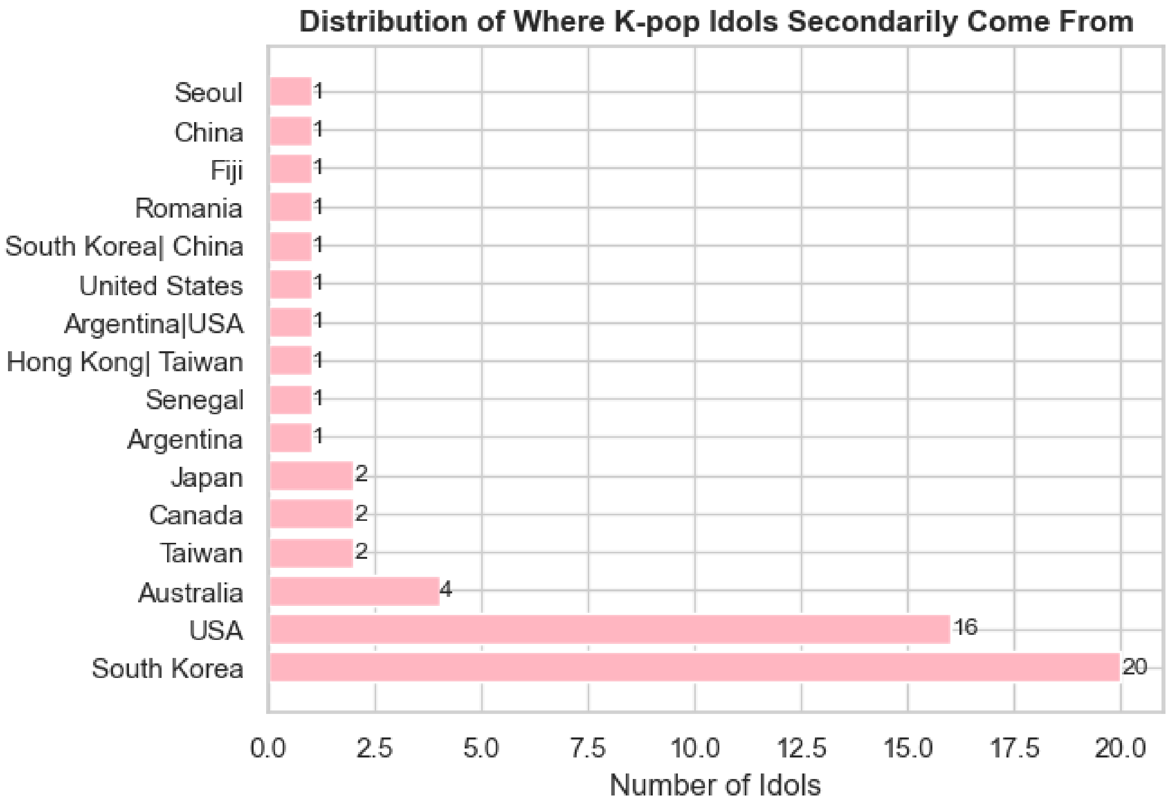
There are also some idols that come from another country on top of their primary country. Let us now take a look at the idols' secondary countries.

In [611…

```
#plotting a horizontal bar graph depicting the secondary countries idols car
#and the distribution of idols in them

#getting the number of idols from the respective secondary countries
idol_countries2 = df_idols['Second Country'].value_counts()

#plotting the horizontal bar graph
plt.barh(idol_countries2.index, idol_countries2, color = 'lightpink')

#adding the labels depicting the exact count of idols from their respective
for index, value in enumerate(idol_countries2):
    plt.text(value, index, str(value), va = 'center', ha = 'left', fontsize

#label and title of graph
plt.xlabel('Number of Idols')
plt.title('Distribution of Where K-pop Idols Secondarily Come From', fontwe:
```

Out[611]:

```
Text(0.5, 1.0, 'Distribution of Where K-pop Idols Secondarily Come From')
```



Based on the graph above, we can see that there are vastly less idols that come from another country. Out of 1604 idols, only 56 came from another country. The leading secondary country is also South Korea, which tells us that even if the idols that did not primarily come from there, they still have some South Korean descent. Our analysis earlier on why this may be the case continues to prove to be true.

Age of Idols Upon Debut:

Aside from descent and nationality, the age of when idols debut is also important. Most of these idols trained for a few years before getting to debut. After debuting, most of them become idols for a long time hence, age is crucial as being an idol is tremendously physically and mentally demanding due to the nature of the job.

From our variable 'idol_stats_count' above, we can see that there are 2 idols that have null values on their date of birth. We can infer this as the count for Date of Birth is 1602, which does not equate to the total number of idols in our research space of 1604. In order to calculate the idols' ages, we would need to remove these anomalies in data.

In [612…

```
#removing the idols with null values on the Date of Birth column
idol_age = df_idols.dropna(subset = ['Date of Birth'])

#making a copy of the data above so we can modify it
idol_age = idol_age.copy()
```

In [613…

```
idol_age.sort_values('Debut')
```

Out[613]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Country	Se Co
1305	Shindong	Shin Donghee	신동희	신동	1985-09-28	Super Junior	2005-11-06	SM	South Korea	
467	Henry	Henry Lau	헨리 라우	헨리	1989-10-11	Super Junior-M	2005-11-06	SM	Canada	T
453	Heechul	Kim Heechul	김희철	희철	1983-07-10	Super Junior	2005-11-06	SM	South Korea	
1643	Yesung	Kim Jongwoon	김종운	예성	1984-08-24	Super Junior	2005-11-06	SM	South Korea	
1425	Sungmin	Lee Sungmin	이성민	성민	1986-01-01	Super Junior	2005-11-06	SM	South Korea	
...	...	...	...	...	...	...	...	...	...	
803	Junghoon	Kim Junghoon	김정훈	정훈	2005-07-05	xikers	2023-03-30	KQ	South Korea	
149	Chi.u	Lee Jaeyi	이재이	치유	1998-08-16	X:IN	2023-04-11	Escrow	South Korea	
1124	Nova	Maria Pavlovna Emelyanova	마리아 파블로브나 에멜야노바	노바	2002-10-10	X:IN	2023-04-11	Escrow	Russia	
278	E.sha	Kwon Yena	권예나	이샤	1998-12-29	X:IN	2023-04-11	Escrow	Australia	
33	Aria	Baby Gauthami	베이비 고타미	아리아	2003-03-12	X:IN	2023-04-11	Escrow	India	

1602 rows × 13 columns

We created a new variable 'idol_age' containing data of idols with valid date values under both the 'Date of Birth' and the 'Debut' columns. We will now calculate and create a new column containing the idols' ages on the day they debuted.

In [614…

```
#calculating the idols' ages on the day of their debut
idol_age.loc[ : ,'Age on Debut'] = -(idol_age['Date of Birth'].dt.year - ido
idol_age.loc[idol_age['Debut'].dt.month < idol_age['Date of Birth'].dt.montl
idol_age.loc[(idol_age['Debut'].dt.month == idol_age['Date of Birth'].dt.mo
             (idol_age['Debut'].dt.day < idol_age['Date of Birth'].dt.day),
```

In [615…

```
#generating a table with sorted values of the 'Debut' in ascending order, i
idol_age.sort_values('Debut')
```

Out[615]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Country	Se Co
1305	Shindong	Shin Donghee	신동희	신동	1985-09-28	Super Junior	2005-11-06	SM	South Korea	
467	Henry	Henry Lau	헨리 라우	헨리	1989-10-11	Super Junior-M	2005-11-06	SM	Canada	T
453	Heechul	Kim Heechul	김희철	희철	1983-07-10	Super Junior	2005-11-06	SM	South Korea	
1643	Yesung	Kim Jongwoon	김종운	예성	1984-08-24	Super Junior	2005-11-06	SM	South Korea	
1425	Sungmin	Lee Sungmin	이성민	성민	1986-01-01	Super Junior	2005-11-06	SM	South Korea	
...	...	...	...	...	...	...	...	...	...	
803	Junghoon	Kim Junghoon	김정훈	정훈	2005-07-05	xikers	2023-03-30	KQ	South Korea	
149	Chi.u	Lee Jaeyi	이재이	치유	1998-08-16	X:IN	2023-04-11	Escrow	South Korea	
1124	Nova	Maria Pavlovna Emelyanova	마리아 파블로브나 에멜야노바	노바	2002-10-10	X:IN	2023-04-11	Escrow	Russia	
278	E.sha	Kwon Yena	권예나	이샤	1998-12-29	X:IN	2023-04-11	Escrow	Australia	
33	Aria	Baby Gauthami	베이비 고타미	아리아	2003-03-12	X:IN	2023-04-11	Escrow	India	

1602 rows × 14 columns

Now that we have determined the idols' age on debut, let us plot this data in a population pyramid. The population pyramid will help us to visualise the age trend of male and female idols upon debuting.

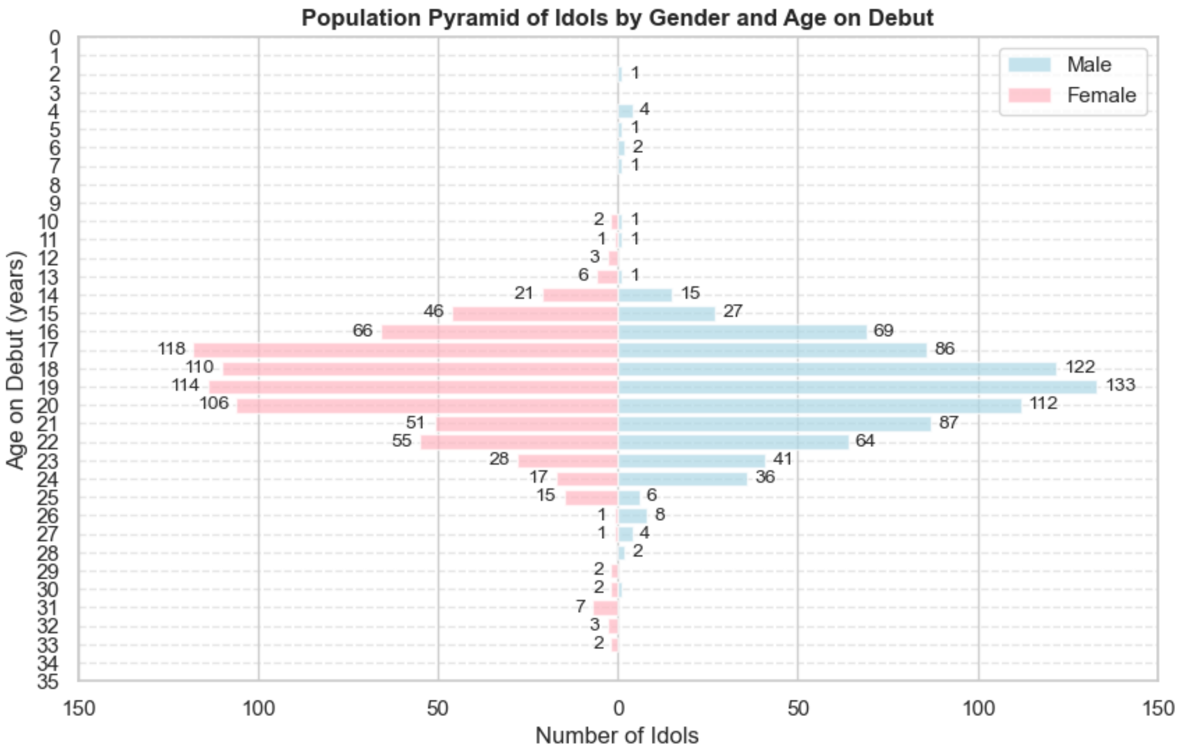
In [616…

```
#separating the males and females
female_data = idol_age[idol_age['Gender'] == 'F']
male_data = idol_age[idol_age['Gender'] == 'M']
```

```
#getting the 'Age on Debut' data for females and males respectively
age_f = pd.Series(female_data['Age on Debut']).value_counts().sort_index()
age_m = pd.Series(male_data['Age on Debut']).value_counts().sort_index()
```

In [617…

```
#plotting of population pyramid
fig, ax = plt.subplots(figsize = (10, 6))
ax.barh(age_m.index, age_m, color = 'lightblue', label = 'Male', alpha = 0.
ax.barh(age_f.index, -age_f, color = 'lightpink', label = 'Female', alpha =

#adding labels for each bar to show the number of individuals in that age ca
for ageM, ageF, count_male, count_female in zip(age_m.index, age_f.index, ag
    ax.text(abs(count_male) + 2, ageM, str(count_male), va = 'center', ha =
    ax.text(-abs(count_female) - 2, ageF, str(count_female), va = 'center',

#labels and title of the population pyramid
ax.set_xlabel('Number of Idols')
ax.set_ylabel('Age on Debut (years)')
ax.set_title('Population Pyramid of Idols by Gender and Age on Debut', fontw

#displays the legend
ax.legend()

#labels each age on the y-axis
ax.set_yticks(range(36))

#forces the all x coordinates to be absolute values
xticks = plt.gca().get_xticks().astype(int)
plt.xticks(xticks, labels = np.abs(xticks));

#inverts y-axis to have younger ages at the top
ax.invert_yaxis()

#adds horizontal lines behind the bars for easier interpretation
ax.grid(axis = 'y', linestyle = '--', alpha = 0.5)

plt.show()
```



From the population pyramid above, we can see that majority of the female idols debut at 17 years' old while majority of the male idols debut at 19 years' old. We can also see that between the ages of 10 - 15 years' old, there were more female idols debuting. Similarly, we can see that between the ages of 23 - 28 years' old, there were more male idols debuting.

In [618…

```
#finding out the average age of the idol when they debut, regardless of gen
mean_age = np.mean(idol_age['Age on Debut'])
print(mean_age)
```

```
19.05680399500624
```

To supplement our research, we find the mean age of debut. From the above, the mean debut age of idols in general is 19 years' old.

Idol Height Analysis:

Up next, we will be analysing one of the more important features: the height. Being the feature that is usually first noticed, the height can be an important factor for these idols as it can add to one's stage presence and image, and heavily impact it. For example, shorter people tend to be seen as cuter while taller people are seen as more charismatic.

In [619…

```
#sorting idols' height in ascending order
df_idols.sort_values('Height')
```

Out[619]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Cou
1048	Mint	Kunpat Phonpawiworakun	쿤팟 퍼언빠위워라쿤	민트	1994-06-23	Tiny-G	2012-08-23	GNG	Thai
240	Dohee	Min Dohee	민도희	도희	1994-09-25	Tiny-G	2012-08-23	GNG	S K
1069	Momoka	Matsuda Momoka	마쓰다 모모카	모모카	2000-12-26	Pink Fantasy	2018-10-28	My Doll	Ja
306	Eunchae	Son Eunchae	손은채	은채	1999-10-06	bugAboo	2021-10-25	A team	S K
1107	Nayoung	Kim Nayoung	김나영	나영	2002-11-30	LIGHTSUM	2021-06-10	Cube	S K
...	...	...	...	...	...	...	...	...	
1766	YY	Kim Moonyong	김문용	와이와이	1991-08-30	UNVS	2020-02-23	CHITWN	S K
1768	Zero	Nasukawa Shota	나스카와 쇼타	제로	2003-01-20	T1419	2007-09-21	CJ E&M	Ja
1771	Zin	Jin Hyunbin	진현빈	지인	2001-08-31	bugAboo	2021-10-25	A team	S K
1775	Zuho	Bae Juho	백주호	주호	1996-07-04	SF9	2016-10-05	FNC	S K
1777	Zuny	Kim Joomi	김주미	주니	1994-12-08	Ladies' Code	2013-03-07	Polaris	S K

1604 rows × 13 columns

From the above table, we see that there are null height values for some idols. This may be likely due to the fact that that information has not been publicly revealed as it is more personal and some people may be sensitive to divulging it. For the sake of our data analysis, we will be removing idols that have null height values.

In [620…

```
#removing idols that have null height values
idol_height = df_idols.dropna(subset = ['Height'])
```

In [621…

```
#data of 5 random idols to check if we managed to remove null height values
idol_height.sample(5)
```

Out[621]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Country	Se Cou
1083	N	Cha Hakyeon	차학연	엔	1990-06-30	VIXX	2012-05-24	Jellyfish	South Korea	
1059	Miso	Jeon Jimin	전지민	미소	2000-10-25	Dreamnote	2018-11-07	iME	South Korea	
508	Hwiseo	Jo Hwihyeon	조휘현	화서	2002-07-31	H1-KEY	2022-01-05	GLG	South Korea	
211	Dasom	Lee Dasom	이다솜	다솜	1993-11-13	2EYES	2013-07-20	SidusHQ	South Korea	
1721	Yue	Nancy Yang	낸시 양	유에	2004-07-03	Lapillus	2022-06-22	MLD	USA	(

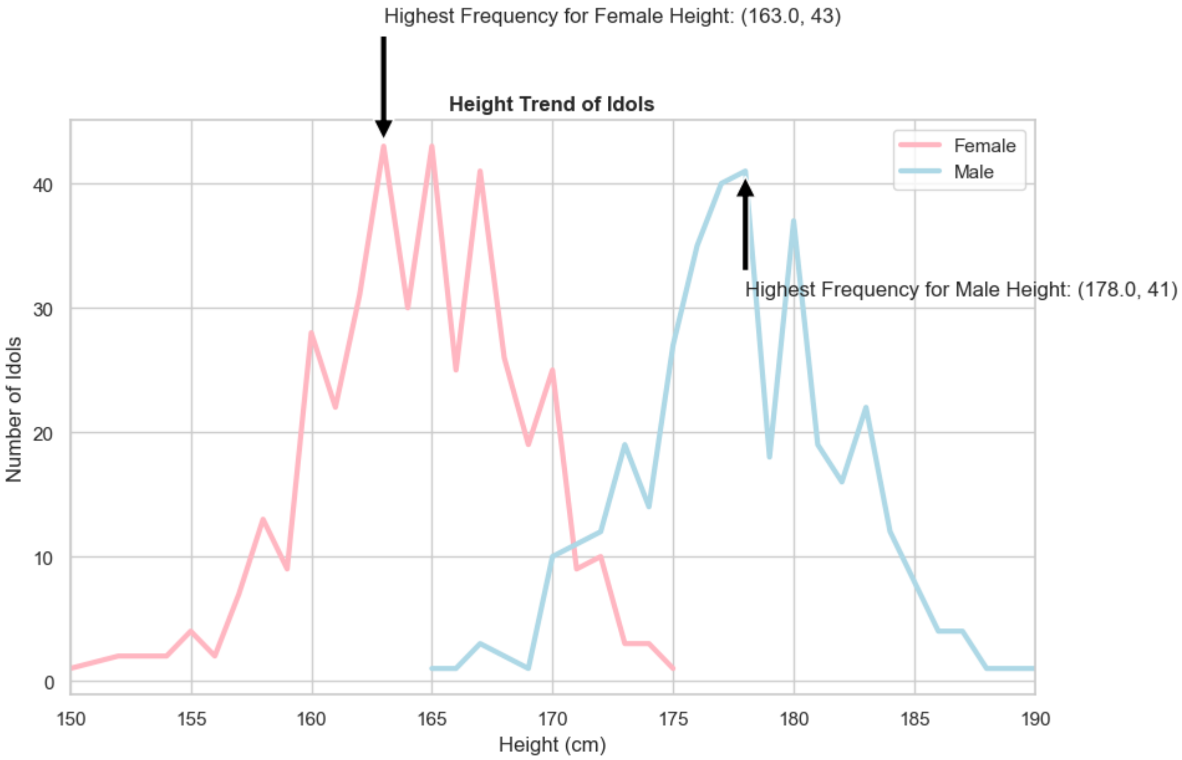
Now that our dataset is more robust, let us visualise this in a lineplot.

In [622…

```
#separating the males and females
female_height = idol_height[idol_height['Gender'] == 'F']
male_height = idol_height[idol_height['Gender'] == 'M']

#getting the 'Height' data for females and males respectively
height_f = pd.Series(female_height['Height']).value_counts().sort_index()
height_m = pd.Series(male_height['Height']).value_counts().sort_index()
```

In [623…

```
#plotting the line graph
plt.figure(figsize=(10, 6))

sns.lineplot(x = height_f.index, y = height_f.values, label = 'Female', colo
sns.lineplot(x = height_m.index, y = height_m.values, label = 'Male', color

#labeling the line graph
plt.xlabel('Height (cm)')
plt.ylabel('Number of Idols')
plt.title('Height Trend of Idols', fontweight = 'bold')

#finding the maximum points, which is the mode height, for both females and
max_height_f = height_f.idxmax()
max_value_f = height_f.max()
max_height_m = height_m.idxmax()
max_value_m = height_m.max()

#adding the labels for mode heights
plt.annotate(f'Highest Frequency for Female Height: ({max_height_f}, {max_va
             xy = (max_height_f, max_value_f),
             xytext = (max_height_f, max_value_f + 10),
             arrowprops = dict(facecolor = 'black', shrink = 0.05))

plt.annotate(f'Highest Frequency for Male Height: ({max_height_m}, {max_valu
             xy = (max_height_m, max_value_m),
             xytext = (max_height_m, max_value_m - 10),
             arrowprops = dict(facecolor = 'black', shrink = 0.05))

#plotting the legend
plt.legend()

#setting the x-axis range starting from the minimum height value
plt.xlim(min(min(height_f.index), min(height_m.index)), max(max(height_f.inc
plt.show()
```



From the above line graph, we can see that male idols are generally taller than female idols as the line graph for male height is towards the higher end of the height spectrum while the line graph for female height is towards the lower end of the height spectrum. For females, the mode height is 163cm with 43 idols while for males, the mode height is 178cm with 41 idols.

```
In [624…
#finding out the average height of the idol, regardless of gender
mean_height = np.mean(idol_height['Height'])
print(mean_height)
```

```
170.75461741424803
```

To supplement our research, we find the mean height of idols. From the above, the mean height of idols in general is 170.6cm.

Idol Weight Analysis:

We will now be analysing one more important feature aside from height- the weight of the individual. Although it is a controversial topic, physical appearance and body shape is heavily focused on in the South Korean entertainment industry. As such, we will be analysing this data as well.

```
In [625…
#sorting idols' weight in ascending order
df_idols.sort_values('Weight')
```

Out[625]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Country
306	Eunchae	Son Eunchae	손은채	은채	1999-10-06	bugAboo	2021-10-25	A team	South Korea
239	Dohee	Kwon Dohee	권도희	도희	2002-08-01	Cignature	2020-02-04	C9	South Korea
240	Dohee	Min Dohee	민도희	도희	1994-09-25	Tiny-G	2012-08-23	GNG	South Korea
1395	Suhye	Kim Suhye	김수혜	수혜	2004-12-13	LIMELIGHT	2023-02-17	143	South Korea
586	Hyunyoung	Cho Hyunyoung	조현영	현영	1991-08-11	Rainbow	2009-11-12	DSP	South Korea
...	...	...	...	...	...	...	...	...	...
1768	Zero	Nasukawa Shota	나스카와 쇼타	제로	2003-01-20	T1419	2007-09-21	CJ E&M	Japan
1771	Zin	Jin Hyunbin	진현빈	지인	2001-08-31	bugAboo	2021-10-25	A team	South Korea
1774	Zoa	Cho Hyewon	조혜원	조아	2005-05-31	Weeekly	2020-07-30	Play M	South Korea
1775	Zuho	Bae Juho	백주호	주호	1996-07-04	SF9	2016-10-05	FNC	South Korea
1777	Zuny	Kim Joomi	김주미	주니	1994-12-08	Ladies' Code	2013-03-07	Polaris	South Korea

1604 rows × 13 columns

From the above table, we see that there are null weight values for some idols. Similar to the height data, this may be likely due to the fact that the information has not been publicly revealed as it is more personal and some people may be sensitive to divulging it. For the sake of our data analysis, we will be removing idols that have null weight values.

In [626…

```
#removing idols that have null weight values
idol_weight = df_idols.dropna(subset = ['Weight'])
```

In [627…

```
#data of 5 random idols to check if we managed to remove null weight values
idol_weight.sample(5)
```

Out[627]:

	Stage Name	Full Name	Korean Name	K Stage Name	Date of Birth	Group	Debut	Company	Country	Se Cou
1296	Seyoung	Ha Seyoung	하세영	세영	1999-02-03	ARTBEAT	2022-11-16	AB Creative	South Korea	
964	Lou	Kim Hosung	김호성	로우	1996-12-21	VAV	2015-10-31	A team	South Korea	
1693	Yoseob	Yang Yoseob	양요섭	요섭	1990-01-05	Highlight	2009-10-14	Around Us	South Korea	
1174	Roda	Shin Joongmin	신중민	로다	1998-09-19	MONT	2014-05-14	Starship	South Korea	
436	Harin	Park Guenhye	박근혜	하린	2000-05-26	Pink Fantasy	2018-10-28	My Doll	South Korea	

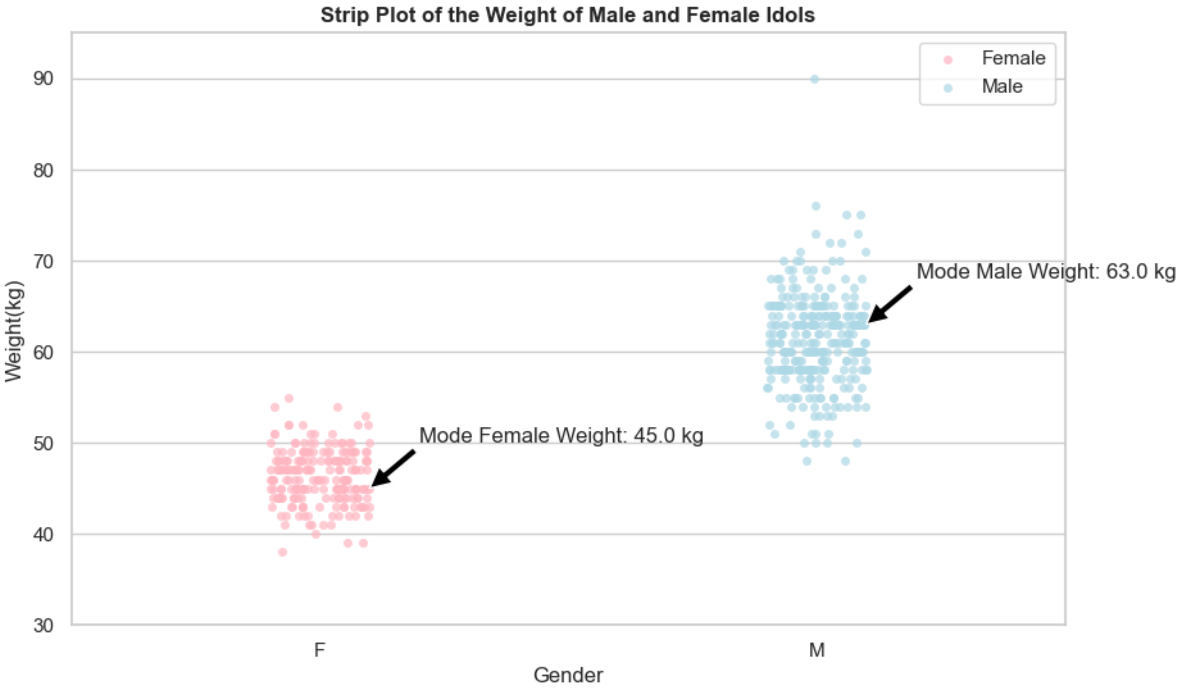
Now that our dataset is more robust, let us visualise our data using a strip plot.

In [628…

```python
#separating the males and females
female_weight = idol_weight[idol_weight['Gender'] == 'F']
male_weight = idol_weight[idol_weight['Gender'] == 'M']

#getting the 'Weight' data for females and males respectively
weight_f = pd.Series(female_weight['Weight']).value_counts().sort_index()
weight_m = pd.Series(male_weight['Weight']).value_counts().sort_index()
```

In [629…

```python
#plotting the strip plot
plt.figure(figsize = (10, 6))

sns.stripplot(x = 'Gender', y = 'Weight', data = female_weight, label = 'Fer
              jitter = True, color = 'lightpink', alpha = 0.7)
sns.stripplot(x = 'Gender', y = 'Weight', data = male_weight, label = 'Male
              jitter = True, color = 'lightblue', alpha = 0.7)

#labeling the plot
plt.xlabel('Gender')
plt.ylabel('Weight(kg)')
plt.title('Strip Plot of the Weight of Male and Female Idols', fontweight =

#finding the mode weight per gender
mode_female = female_weight['Weight'].mode().iloc[0]
mode_male = male_weight['Weight'].mode().iloc[0]

#adding labels for the mode weights
plt.annotate(f'Mode Female Weight: {mode_female} kg', xy = (0.1, mode_female
             arrowprops = dict(facecolor = 'black', shrinkA = 0.05, shrinkB

plt.annotate(f'Mode Male Weight: {mode_male} kg', xy = (1.1, mode_male), xy
             arrowprops = dict(facecolor = 'black', shrinkA = 0.05, shrinkB

#setting the range of values for the y axis
plt.ylim(30, 95)

#plotting the legend
plt.legend()

plt.show()
```



From the strip plot above, we can see that male idols are mainly heavier than female idols because most of the plots are towards the higher end of the weight spectrum while the plots for female weight is towards the lower end. For females, the mode weight is 45kg while for males, the mode weight is 63kg. The difference in both mode weights correlate to the mode heights- while the mode height of males is taller than females, the mode weight of males is also heavier than females.

```
In [630…
#finding out the average weight of the idol, regardless of gender
mean_weight = np.mean(idol_weight['Weight'])
print(mean_weight)
```

```
57.437137330754354
```

To supplement our research, we find the mean weight of idols. From the above, the mean weight of idols in general is 57.4kg.

Agency Company Analysis:

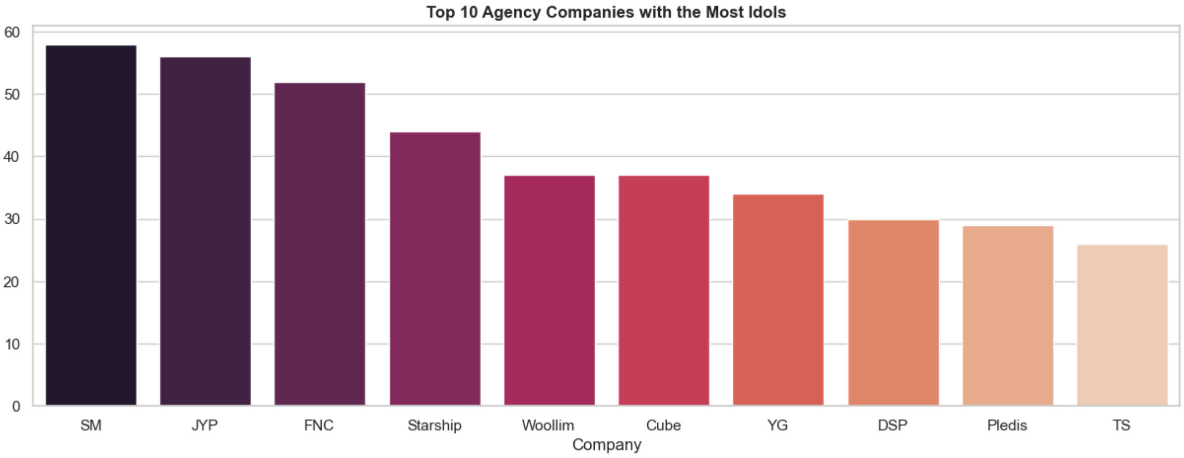
We will now analyse the agency companies that the idols are in. Aside from certain physical characteristics, the way individuals are trained and assessed may affect their chances of being a K-pop idol. These companies are integral in the growth and development of an individual- starting from their trainee days, up until they finish their careers as a debuted idol. We will first find out how many different companies there are in the dataset.

```
In [631…
#finding out how many different companies are in the dataset
number_of_companies = df_idols['Company'].nunique()
print(number_of_companies)
```

```
176
```

For the sake of our aim, we would like to find what may be the best company to be in in order to debut to be a successful K-pop idol. To achieve this, we will be narrowing the scope down and only be focusing on the top 10 companies.

```
In [632…  #making a bar graph depicting the top 10 companies
          plt.figure(figsize = (15, 5))
          sns.barplot(x = df_idols['Company'].value_counts().head(10).index,
                      y = df_idols['Company'].value_counts().head(10).values, palette

          plt.title('Top 10 Agency Companies with the Most Idols', fontweight = 'bold
          plt.show()
```



From the bar graph above, we can see that the top company with the most debuted idols over the past 20 years is SM, followed by JYP, and then FNC. The company debuting the most idols may mean that there are higher chances of an individual debuting, meaning that trainees would have had sufficient and rigorous training to be deemed ready to debut. Hence, aside from having certain individual characteristics, being in SM may play a part in being a successful K-pop idol. However, this may not always be true as we are only looking at quantitative data. Although SM has the most debuted idols, they may not have the best training quality. Other companies that have less idols may have better quality teaching.

Conclusion

Through our thorough technical analysis of the dataset above, we can make a conclusion and have the answer to what may be the optimum traits an individual may have in order to debut to be a K-pop idol.

The optimum age to debut is 19 years' old for males and 17 years' old for females. However, it is worth noting that this does not take into account the training duration. Some idols train for as short as a few months while some train for as long as 10 years. Hence, the optimum age for an individual to start training to be a K-pop idol would be below 19 and 17 years' old for males and females respectively.

We also found out that it serves as a big advantage to be of South Korean descent, or similar East Asian descent- regardless if the individual is only partially or fully blooded. As K-pop's primary audience are the South Koreans, it serves as a huge advantage to know their language. Aside from that, the more obvious physical traits that may be needed to cater to the aforementioned audience may be body build, hair type and skin colour. South Koreans, and East Asians in general, are stereotypically more fair-skinned, straight-haired, and for females: more on the petite side.

Speaking of physical appearances, the optimum height for females is 163cm with a weight of 45.0kg. For males, the optimum height is 178cm with a weight of 63.0kg. In general, individuals with a height of 170.6cm and a weight of 57.4kg is optimal.

Lastly, the agency company one may be in in order to have the highest chance of debuting would be SM. The company an individual is under will greatly influence their growth and development, and ultimately their chances of being a K-pop idol.

All in all, the characteristics above are considered optimum based on the analysis of the data of idols who have debuted in the past 20 years. However, this barely serves as any indication on whether or not an individual can become a K-pop idol. There are many more factors that go into a company's decision to debut a trainee- such as singing and dancing abilities, all-roundedness, and stage presence.

Reference

(NNK), N. (2024, January 9). How to format pandas Datetime?. Spark By {Examples}. https://sparkbyexamples.com/pandas/how-to-format-pandas-datetime/#google_vignette

Alayo, N. (2023, May 5). 1700+ K-pop idols dataset. Kaggle. https://www.kaggle.com/datasets/nicolsalayoarias/all-kpop-idols/data

Creative commons, CC0 1.0 Deed. CC0 1.0 Deed | CC0 1.0 Universal | Creative Commons. (n.d.). https://creativecommons.org/publicdomain/zero/1.0/

Hankyoreh. (n.d.). How Lee Soo-man’s idol system at SM paved the way for K-pop as we know it. https://english.hani.co.kr/arti/english_edition/e_entertainment/1084898.html

Mukhtiar, M. (2023, August 14). Visualizing Big Data: Techniques for handling and displaying large data sets. Medium. https://medium.com/plumbersofdatascience/visualizing-big-data-techniques-for-handling-and-displaying-large-data-sets-84839c493202#:~:text=Hierarchical%20Visualization%3A%20Hierarchical%20visualizatior

Nguyen, X. (2021, January 22). Data visualization: How to choose the Right Chart [part 2]. Medium. https://towardsdatascience.com/data-visualization-how-to-choose-the-right-chart-part-2-fb32ed14c7be

Statistical Data Visualization. seaborn. (n.d.). https://seaborn.pydata.org/index.html

What percentage of people who audition actually become K-pop trainees?. Quora. (n.d.). https://www.quora.com/What-percentage-of-people-who-audition-actually-become-K-pop-trainees