Determinants and Effects of Admission to Selective Private Colleges

Econ 148

Professor Tamer Cetin

Student Group 7:

Saga Abuodeh, Anne Al-Azzawi, Kashvi Patel, Kashyap Coimbatore Murali, Yikang Yu, Yosolaoluwa Olakunle

Project Background

The journal article **Diversifying Society's Leaders? The Determinants and Causal Effects of Admission to Highly Selective Private Colleges** by Chetty et al. investigates the factors influencing college admissions, particularly at highly selective private colleges, and their implications for socioeconomic diversity and post-college outcomes.

The study examines how academic qualifications, non-academic factors, and parental income **affect admissions rates** and **post-college success.** It also discusses the impact of legacy preferences, athletic recruitment, and holistic evaluation policies on **admissions advantages for high-income families.** Additionally, the article explores the potential for policy changes to increase economic mobility and reduce disparities in admissions rates based on parental income.

In this study, the **data used** came from 4 sources:

- Federal income tax records, 1996-2021
- Federal college attendance records, 1999-2015
- Standardized test score data, 2001-2015
- Applications and Admissions Records from Colleges
  - Several Ivy-Plus colleges, various years
  - Highly Selective Public institutions: UC-Berkeley, UCLA, UT-Austin, plus other most selective public flagships
  - Several college systems: UC system, Cal State, Texas system (THECB)
  - Detailed student characteristics, admissions outcomes, internal rating

Project Background

Article’s Key Claims:

1. Financial resources alone can't improve economic opportunity.

2. Holistic policies benefit high-income students in admissions.

3. **Non-academic credentials give high-income families an edge.**

4. **Academic qualifications predict success; non-academic factors don't.**

5. Selective colleges reinforce privilege but can change.

Project Objective

Expanding on Chetty et al.'s paper, we further explored specific subquestions:

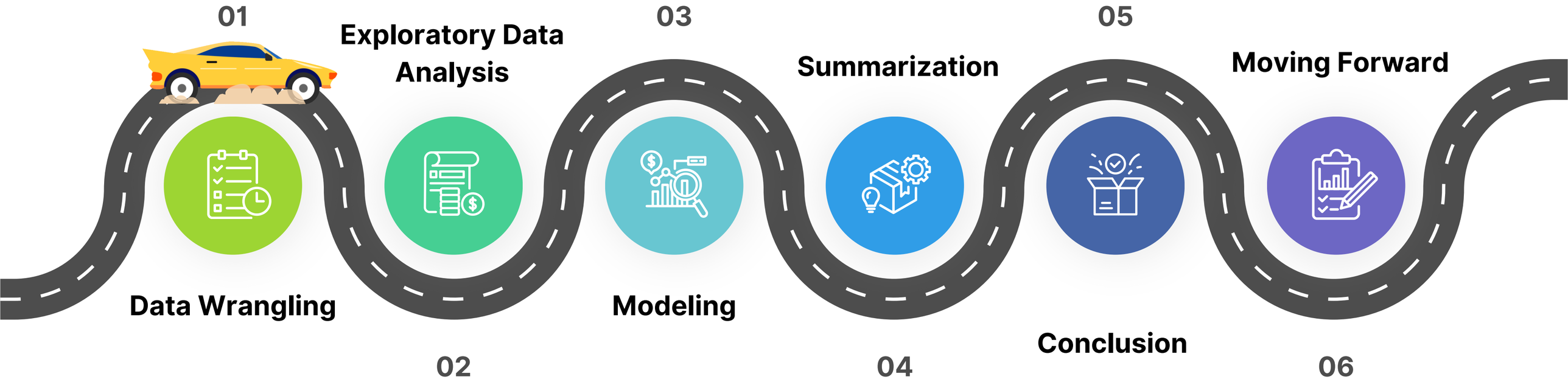
The purpose of this project is to investigate how coming from a high-income family influences a student's chances of being admitted to an Ivy-Plus college, exploring the factors contributing to these admissions. Additionally, we aim to examine the impact of SAT and ACT submissions on admission rates, predicting the admission rates into Ivy-Plus and other elite colleges based on these scores. We will employ **predictive models that integrate financial and academic indicators** to analyze these dynamics.

Key Research Questions:

01. Does coming from a high-income family affect a student's chances of being admitted to an Ivy-Plus college, and what factors contribute to this?

02. Do SAT and ACT submissions affect admission rates? Predict the admission rates into Ivy-Plus and other elite colleges based on SAT/ACT scores

Roadmap Research Process



The Data

Origin: Opportunity Insights Date Base

Dimension: 1713 observations, 81 variables

Structure: Each row represents a unique applicant, while each column represents a different attribute or variable related to college admissions and socioeconomic factors.

Granularity: The granularity of the dataset is at the institutional level, meaning each row provides data for one applicant at one institution. The columns include various metrics such as attendance rates, parental income brackets, SAT scores, and institutional characteristics, etc.

Scope: The scope of the dataset includes a wide range of variables that capture different aspects of college admissions and socioeconomic factors. This includes detailed information on attendance rates, application rates, parental income, and institutional characteristics.

Temporality: Cross-sectional dataset, capturing data from a specific period. The exact time frame is not specified,represents snapshot of college admissions 1998-2015

Faithfulness: Captures various dimensions of college admissions accurately. However, there may be limitations in terms of missing values and the level of detail for certain variables.

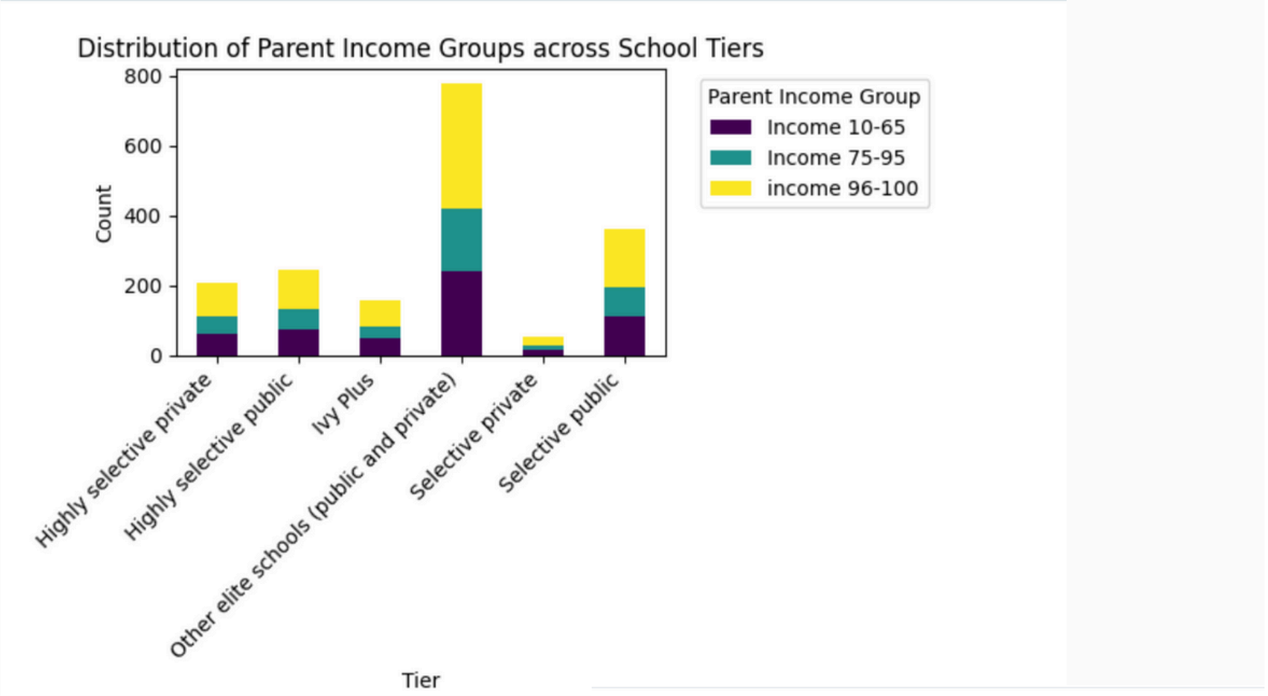
CollegeAdmissions_Data.csv

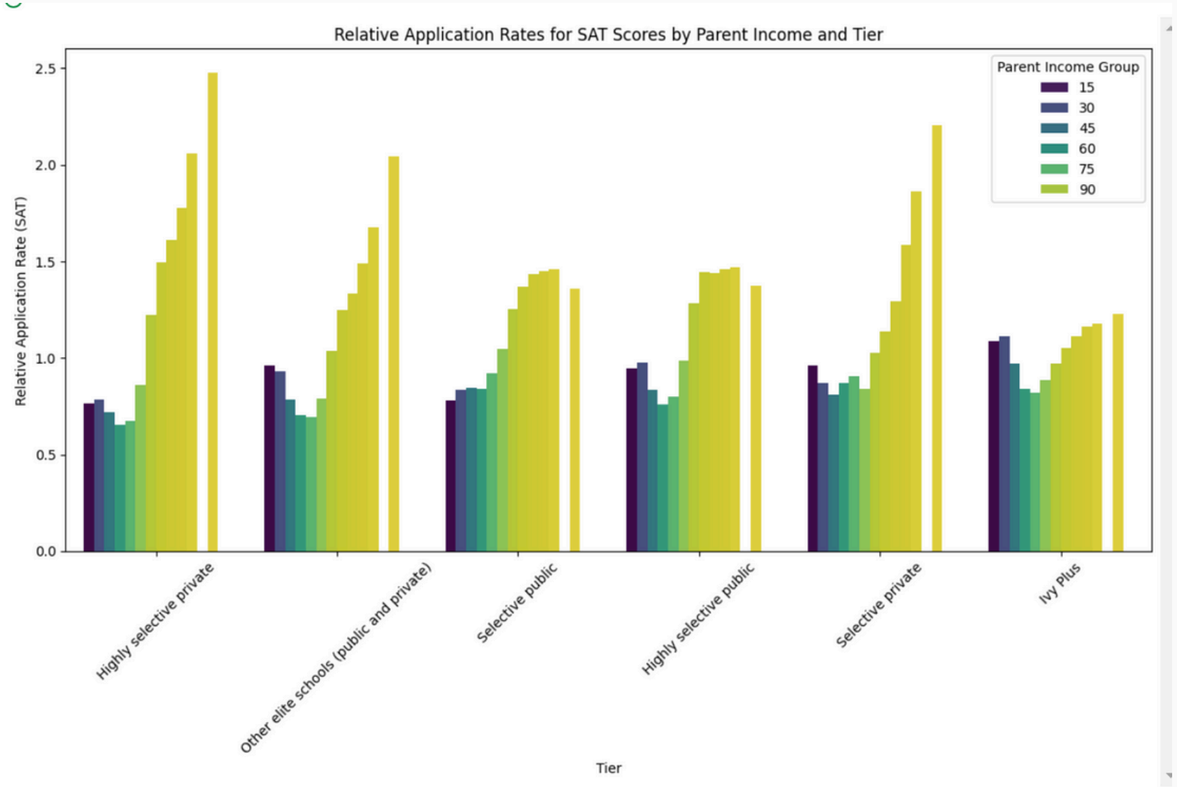
super_opeid object	name object	par_income_bin ob...	par_income_lab ob...
1434	American Univers...	10	0-20
1434	American Univers...	30	20-40
1434	American Univers...	50	40-60
1434	American Univers...	65	60-70
1434	American Univers...	75	70-80
1434	American Univers...	85	80-90
1434	American Univers...	92.5	90-95
1434	American Univers...	95.5	95-96
1434	American Univers...	96.5	96-97
1434	American Univers...	97.5	97-98

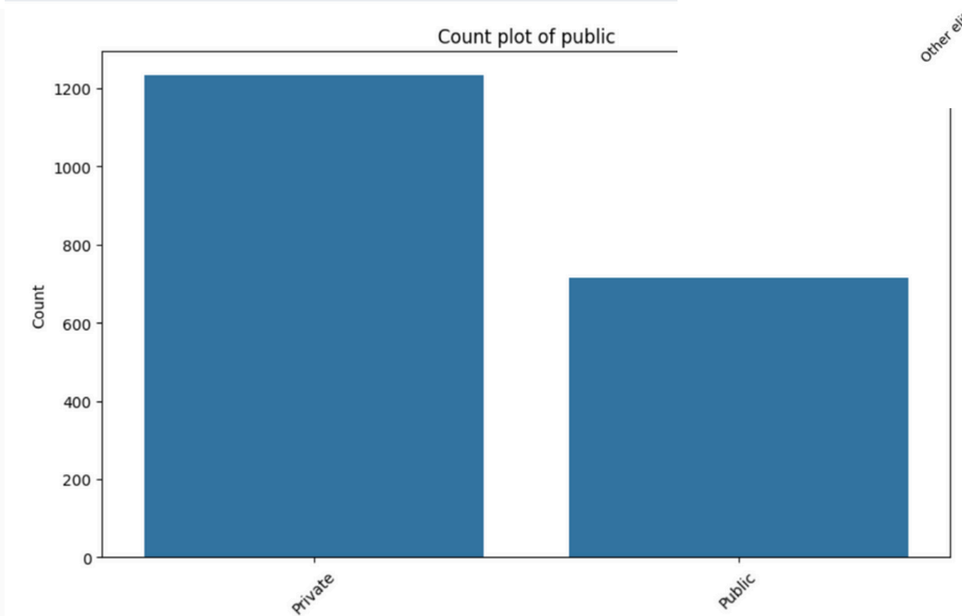
1713 rows, 81 cols, showing 10 rows/page

***Since we did not have admissions rate for any college we used attendance rate as a proxy . Low admission rate usually would indicate high attendance etc.**

The Data


Distribution of Parent Income Groups across School Tiers
Parent Income Group
Income 10-65
Income 75-95
income 96-100
Count
800
600
400
200
0
Highly selective private
Highly selective public
Ivy Plus
Other elite schools (public and private)
Selective private
Selective public
Tier


Relative Application Rates for SAT Scores by Parent Income and Tier
Parent Income Group
15
30
45
60
75
90
Relative Application Rate (SAT)
2.5
2.0
1.5
1.0
0.5
0.0
Highly selective private
Other elite schools (public and private)
Selective public
Highly selective public
Selective private
Ivy Plus
Tier


Count plot of public
Count
1200
1000
800
600
400
200
0
Private
Public

The Data

Based on our expertise and insights from the variable code book/legend, we **hand selected specific variables** that offer relevant information to **address the questions at hand**.
We ended up with 15 variables in total.

Variable Description:

name: Name of college (or college group)
par_income_bin: Parent household income group (and the label) based on percentile in the income distribution.
rel_apply: Test-score-reweighted relative application rate.
rel_attend: Test-score-reweighted absolute and relative
attendance rate:
rel_apply_sat: Relative application rate for specific test score band based on school tier/category
attend_instate: Test-score-reweighted absolute and relative
attendance rate for in-state students.
attend_instate_sat: Absolute and relative estimates on a specific testscore for in-state students.
attend_oostate: Test-score-reweighted absolute and relative
attendance rate for out-of-state students.
attend_oostate_sat: Absolute and relative estimates on a specific testscore for out-of-state students.
attend_unwgt:Unweighted absolute and relative attendance rate.
attend_unwgt_instate: Unweighted absolute and relative estimates for instate students.
attend_unwgt_oostate Unweighted absolute and relative estimates for out-of-state students.
public: Indicator for public universities
tier_name: Selectivity and type combination:
1 = Ivy-Plus (Ivy League colleges plus Stanford, Chicago, Duke, and MIT)
2 = Other elite college (Barron's top selectivity category, other than the Ivy-plus, both public and private combined)
3 = Highly selective public college (Barron's 2nd selectivity group)
4 = Highly selective private college (Barron's 2nd selectivity group)
5 = Selective public college (Barron's 3rd, 4th, and 5th selectivity groups)
6 = Selective private college (Barron's 3rd, 4th, and 5th selectivity groups)
See Chetty, Friedman, Saez, Turner, and Yagan (2020) for more information on how the tier is defined

```
columns_list = [
    'name',
    'par_income_bin',
    'rel_apply',
    'tier',
    'rel_attend',
    'rel_apply_sat',
    'attend_instate',
    'attend_instate_sat',
    'attend_oostate',
    'attend_oostate_sat',
    'attend_unwgt',
    'attend_unwgt_instate',
    'attend_unwgt_oostate',
    'public', 'tier_name'
]

df_cleaned = df[columns_list ]
df_cleaned
```

	name object	par_income_bin flo...	rel_apply float64	tier object	rel_attend float64	rel_apply_sat float
	American Un... 0.7% Amherst Coll... 0.7% 137 others 98.6%	10.0 - 100.0	0.070443429 - 5.8...	Other elite ... 43.2% Selective pu... 20.1% 4 others 36.7%	0.010013658 - 10.2...	0.18820071 - 4.70
20	Amherst College	92.5	0.84161502	Other elite school...	0.63088179	0.895215
21	Amherst College	95.5	0.94007629	Other elite school...	1.0127096	0.805948
22	Amherst College	96.5	1.0177214	Other elite school...	0.81828016	1.09247
23	Amherst College	97.5	1.0992246	Other elite school...	1.0369351	1.25003
24	Amherst College	98.5	1.3969384	Other elite school...	1.7949971	1.61727
25	Amherst College	99.400002	1.5993084	Other elite school...	1.9063946	r
26	Amherst College	99.5	1.617566	Other elite school...	2.051353	1.93576
27	Amherst College	100	1.7539662	Other elite school...	2.9204836	r
28	Auburn University	10	0.75778812	Selective public	0.56999844	0.754524
29	Auburn University	30	0.82170606	Selective public	0.62039053	0.744011

The Data

Based on our expertise and insights from the variable code book/legend, we **hand selected specific variables** that offer relevant information to **address the questions at hand**.
We ended up with 15 variables in total.

Variable Description:

name: Name of college (or college group)
par_income_bin: Parent household income group (and the label) based on percentile in the income distribution.
rel_apply: Test-score-reweighted relative application rate.
rel_attend: Test-score-reweighted absolute and relative
attendance rate:
rel_apply_sat: Relative application rate for specific test score band based on school tier/category
attend_instate: Test-score-reweighted absolute and relative
attendance rate for in-state students.
attend_instate_sat: Absolute and relative estimates on a specific testscore for in-state students.
attend_oostate: Test-score-reweighted absolute and relative
attendance rate for out-of-state students.
attend_oostate_sat: Absolute and relative estimates on a specific testscore for out-of-state students.
attend_unwgt:Unweighted absolute and relative attendance rate.
attend_unwgt_instate: Unweighted absolute and relative estimates for instate students.
attend_unwgt_oostate Unweighted absolute and relative estimates for out-of-state students.
public: Indicator for public universities
tier_name: Selectivity and type combination:
1 = Ivy-Plus (Ivy League colleges plus Stanford, Chicago, Duke, and MIT)
2 = Other elite college (Barron's top selectivity category, other than the Ivy-plus, both public and private combined)
3 = Highly selective public college (Barron's 2nd selectivity group)
4 = Highly selective private college (Barron's 2nd selectivity group)
5 = Selective public college (Barron's 3rd, 4th, and 5th selectivity groups)
6 = Selective private college (Barron's 3rd, 4th, and 5th selectivity groups)
See Chetty, Friedman, Saez, Turner, and Yagan (2020) for more information on how the tier is defined

```
columns_list = [
    'name',
    'par_income_bin',
    'rel_apply',
    'tier',
    'rel_attend',
    'rel_apply_sat',
    'attend_instate',
    'attend_instate_sat',
    'attend_oostate',
    'attend_oostate_sat',
    'attend_unwgt',
    'attend_unwgt_instate',
    'attend_unwgt_oostate',
    'public', 'tier_name'
]

df_cleaned = df[columns_list ]
df_cleaned
```

	name object	par_income_bin flo...	rel_apply float64	tier object	rel_attend float64	rel_apply_sat float
	American Un... 0.7% Amherst Coll... 0.7% 137 others 98.6%	10.0 - 100.0	0.070443429 - 5.8...	Other elite ... 43.2% Selective pu... 20.1% 4 others 36.7%	0.010013658 - 10.2...	0.18820071 - 4.70
20	Amherst College	92.5	0.84161502	Other elite school...	0.63088179	0.895215
21	Amherst College	95.5	0.94007629	Other elite school...	1.0127096	0.805948
22	Amherst College	96.5	1.0177214	Other elite school...	0.81828016	1.09247
23	Amherst College	97.5	1.0992246	Other elite school...	1.0369351	1.25003
24	Amherst College	98.5	1.3969384	Other elite school...	1.7949971	1.61727
25	Amherst College	99.400002	1.5993084	Other elite school...	1.9063946	r
26	Amherst College	99.5	1.617566	Other elite school...	2.051353	1.93576
27	Amherst College	100	1.7539662	Other elite school...	2.9204836	r
28	Auburn University	10	0.75778812	Selective public	0.56999844	0.754524
29	Auburn University	30	0.82170606	Selective public	0.62039053	0.744011

The Data

As with all real world data, our data had **missing values**. We encountered a challenge at this stage. To address it, we explored two approaches: **(1)** imputing the missing data, and **(2)** dropping columns with more than 30% missing data. We ended up using both approaches and compared.

Tradeoffs

- **Imputing too much** data can result in a lack of natural variation, making the **model ineffective**.
- Dropping up to 6 columns reduces our available variables, thereby **limiting the number of features** we can use for modeling. This reduction may **affect** the **model's ability** to capture nuanced relationships and potentially lead to **less robust** predictions or analyses.

***Most of the missing data is from private school data.**

Nan Count

name	0
par_income_bin	0
rel_apply	0
tier	0
rel_attend	2
rel_apply_sat	278
attend_instate	1334
attend_instate_sat	1334
attend_oostate	1336
attend_oostate_sat	1346
attend_unwgt	1
attend_unwgt_instate	1334
attend_unwgt_oostate	1337
public	0
tier_name	0

dtype: int64

Nan %

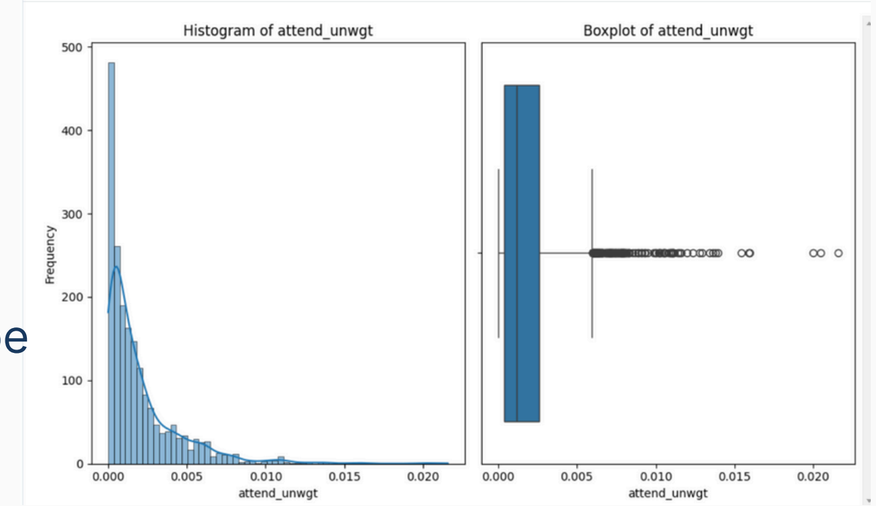
name	0.000000
par_income_bin	0.000000
rel_apply	0.000000
tier	0.000000
rel_attend	0.102775
rel_apply_sat	14.285714
attend_instate	68.550874
attend_instate_sat	68.550874
attend_oostate	68.653649
attend_oostate_sat	69.167523
attend_unwgt	0.051387
attend_unwgt_instate	68.550874
attend_unwgt_oostate	68.705036
public	0.000000
tier_name	0.000000

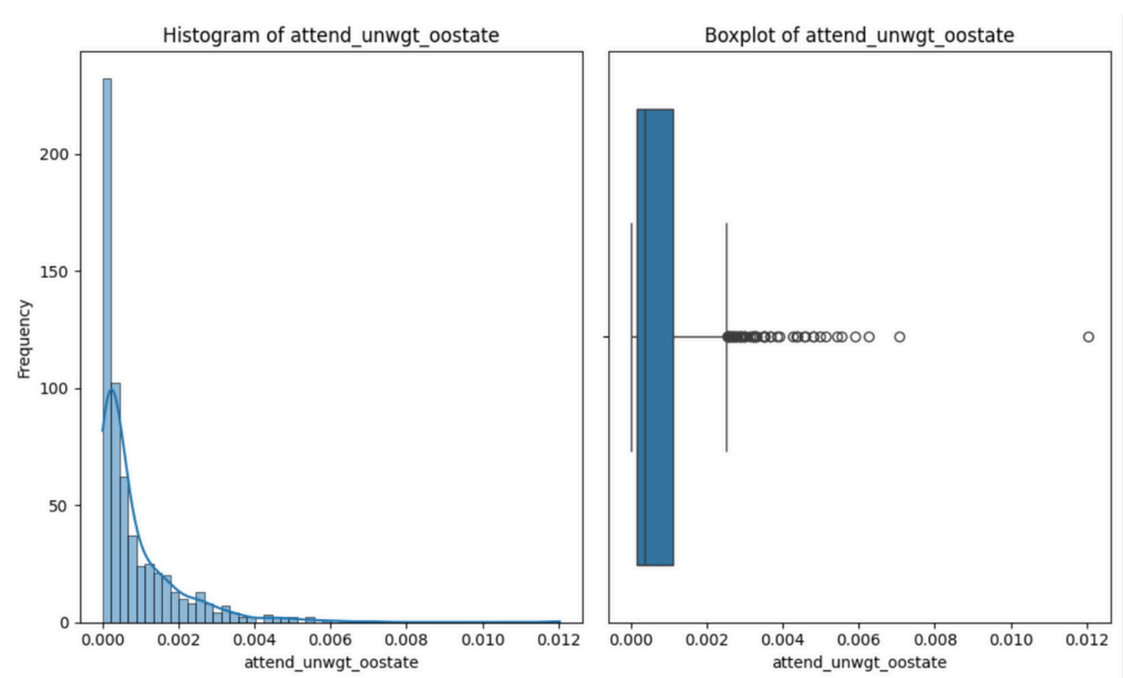
dtype: float64

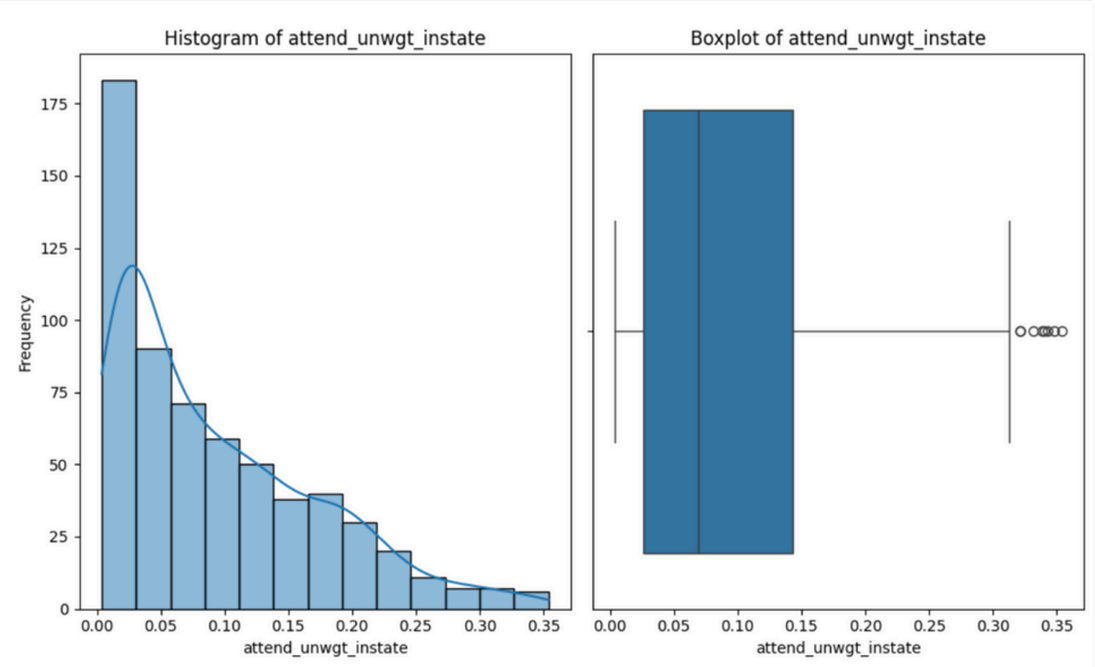
The Data

Imputation

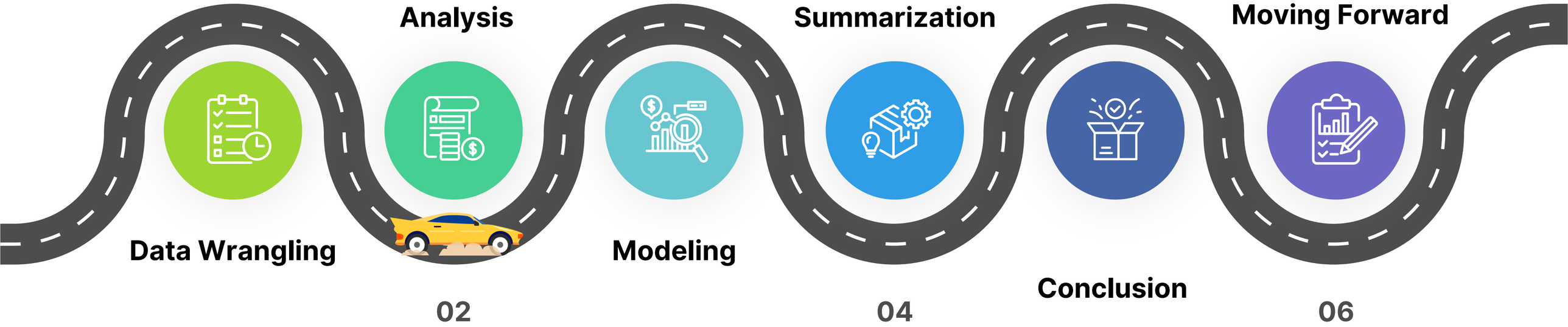
- For imputation, we decided to use the median since when the **data is skewed the median is more useful**. The mean would be distorted by outliers and the median would maintain the distributions' characteristics.







Roadmap Research Process



EDA

Variable Statistics

```
1 stats_impute = df_cleaned_imputed.describe()
2 stats_impute
```

	par_income_bin flo...	rel_apply float64	rel_attend float64	rel_apply_sat float64	attend_instate floa...	attend_instate_sat f..	attend_oostat
cou...	1946	1946	1946	1946	1946	1946	
mean	78.17142871	1.172632366	1.272998122	1.133327855	0.127348838	0.1135597707	0.00093077
std	28.04041119	0.5611611184	0.9246311507	0.4732076745	0.04751472323	0.05826043113	0.00099270
min	10	0.070443429	0.010013658	0.18820071	0.0042952062	0.0016169919	0.000005
25%	65	0.8319239175	0.759556045	0.8297781925	0.12593888	0.10660941	0.0007361
50%	94	1.048234	1.02585195	1.0560932	0.12593888	0.10660941	0.0007361
75%	98.5	1.3367169	1.4464752	1.287862575	0.12593888	0.10660941	0.0007361
max	100	5.8736005	10.261024	4.7022381	0.41473576	0.65674222	0.0191

8 rows, 11 cols, showing 10 rows/page Page 1 of 1

```
1 stats_dropped = df_cleaned_dropped.describe()
2 stats_dropped
```

	par_income_bin flo...	rel_apply float64	rel_attend float64	rel_apply_sat float64	attend_unwgt float...
cou...	1946	1946	1946	1946	1946
mean	78.17142871	1.172632366	1.272998122	1.133327855	0.002001481343
std	28.04041119	0.5611611184	0.9246311507	0.4732076745	0.002427203064
min	10	0.070443429	0.010013658	0.18820071	0.0000017983402
25%	65	0.8319239175	0.759556045	0.8297781925	0.000366359545
50%	94	1.048234	1.02585195	1.0560932	0.0011591336
75%	98.5	1.3367169	1.4464752	1.287862575	0.00261328065
max	100	5.8736005	10.261024	4.7022381	0.021571836

8 rows, 5 cols, showing 10 rows/page Page 1 of 1

```
1 variances_imputed
```

```
par_income_bin          7.862647e+02
rel_apply               3.149018e-01
rel_attend              8.549428e-01
rel_apply_sat           2.239255e-01
attend_instate          2.257649e-03
attend_instate_sat      3.394278e-03
attend_oostate          9.854655e-07
attend_oostate_sat      6.398377e-07
attend_unwgt            5.891315e-06
attend_unwgt_instate    2.096936e-03
attend_unwgt_oostate    4.678375e-07
dtype: float64
```

```
1 variances_dropped
```

```
par_income_bin      786.264660
rel_apply             0.314902
rel_attend            0.854943
rel_apply_sat         0.223926
attend_unwgt          0.000006
dtype: float64
```

EDA (Correlation)

Correlation

- After calculating the variance, we then meaured the correlation between rel_apply and rel_attend variables, we used a heat map to visulaize our findings.

High Positive Correlations:

rel_apply and rel_attend (0.92): These variables are highly positively correlated, suggesting that students who apply are very likely
This indicates a strong relationship between the application process and actual attendance.

Moderate Positive Correlations:

rel_apply and rel_apply_sat (0.65): Students who apply are moderately likely to have higher SAT scores. This suggests that the like
applying is somewhat related to SAT scores.

rel_apply and attend_unwgt (0.43): There is a moderate positive correlation between application rates and unweighted attendance
indicates that students who apply are moderately likely to attend, irrespective of weighted factors.

rel_attend and rel_apply_sat (0.52): There is a moderate positive correlation between attendance and SAT scores among applicant
that higher SAT scores among applicants are somewhat related to actual attendance.

rel_attend and attend_unwgt (0.49): This shows a moderate correlation between attendance and unweighted attendance, suggest
relationship between relative attendance rates and overall attendance.

Low Positive Correlations:

par_income_bin and rel_apply (0.39): Parental income has a moderate influence on students' likelihood to apply. This suggests tha
parental income can slightly increase application rates.

par_income_bin and rel_attend (0.36): Parental income also moderately influences attendance rates, indicating that higher income
students who are more likely to attend.

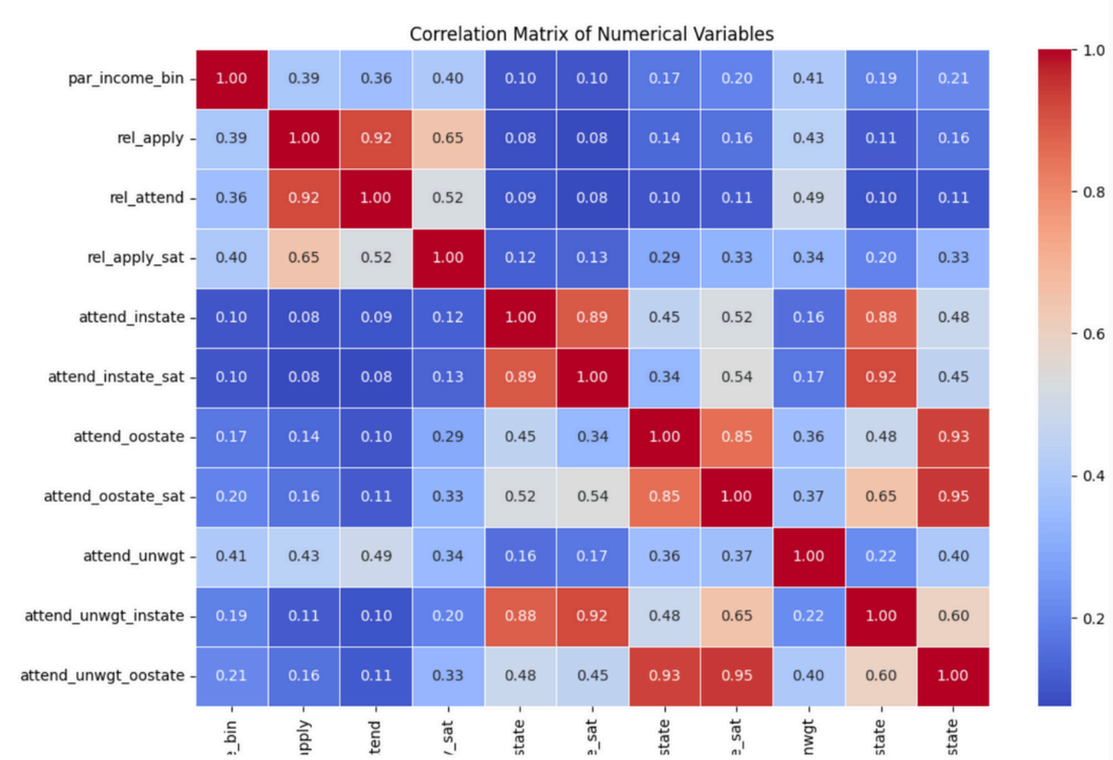
par_income_bin and rel_apply_sat (0.40): There is a moderate positive correlation between parental income and SAT scores among
suggesting that students from higher-income families tend to have higher SAT scores.

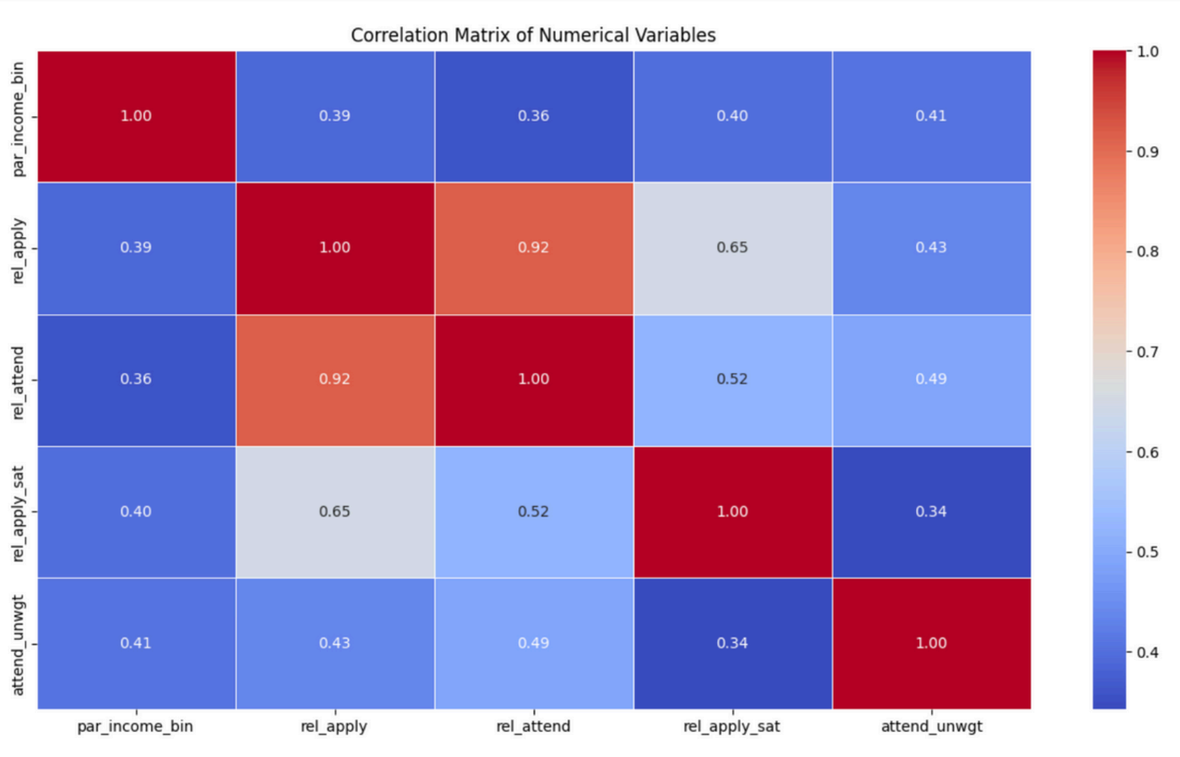
par_income_bin and attend_unwgt (0.41): Parental income has a moderate influence on unweighted attendance, implying that stud
higher-income families are more likely to attend.

rel_apply_sat and attend_unwgt (0.34): There is a low to moderate positive correlation between SAT scores among applicants and
attendance, suggesting a relationship between higher SAT scores and attendance rates.

EDA (Correlation) Cont.

Heat Map Correlation





EDA (Correlation) Cont.

High Correlations Indicating Potential Multicollinearity:

High Correlations Indicating Potential Multicollinearity:

rel_apply and rel_attend (0.92): These variables are highly correlated, suggesting that one could be a proxy for the other in a regression model

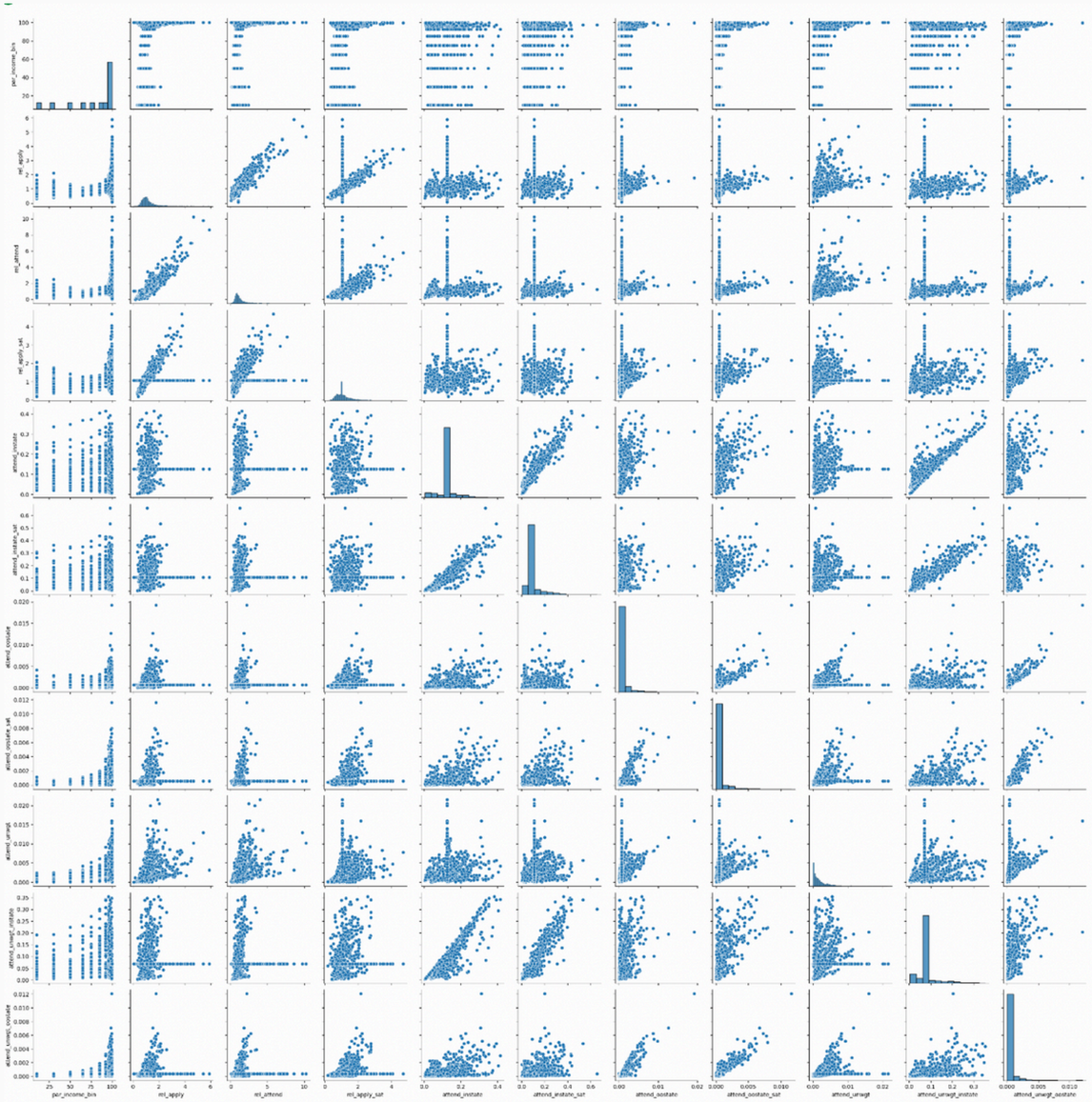
attend_instate and attend_instate_sat (0.90): Strong correlation indicates potential redundancy.

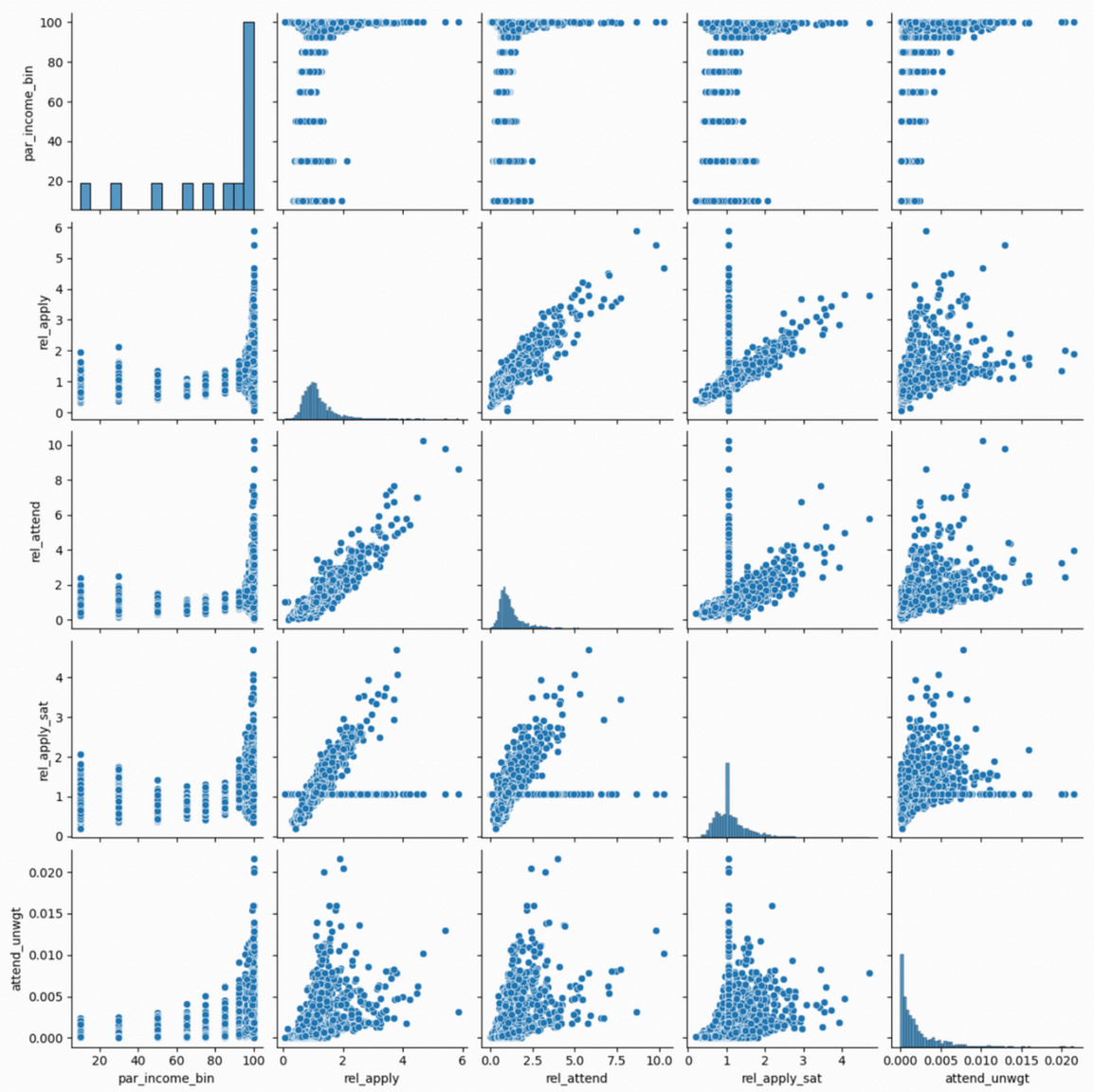
attend_unwgt_instate and attend_instate_sat (0.92): High correlation suggests multicollinearity concerns.

attend_oostate and attend_oostate_sat (0.83): Moderate to high correlation indicates potential issues.

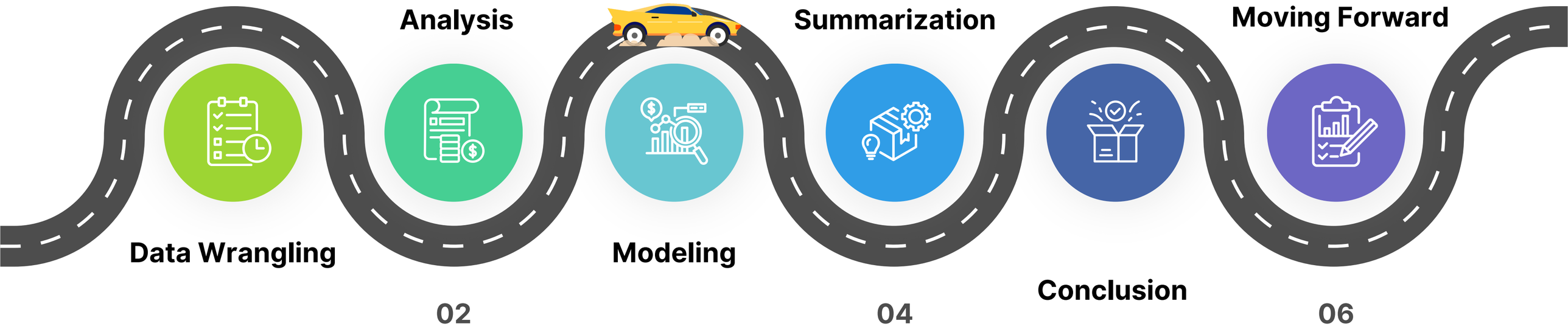
attend_unwgt_oostate and attend_oostate_sat (0.94): Very high correlation indicates that including both in a regression model might cause multicollinearity.

EDA (Pairplot) Graphs





Roadmap Research Process



Modeling Q1: Logit Regression Results

Socio-economic and application-related variables — Logit

Dep. Variable:	is_ivy	No. Observations:	1556
Model:	Logit	Df Residuals:	1552
Method:	MLE	Df Model:	3
Date:	Thu, 27 Jun 2024	Pseudo R-squ.:	0.02105
Time:	04:09:47	Log-Likelihood:	-437.70
converged:	True	LL-Null:	-447.12
Covariance Type:	nonrobust	LLR p-value:	0.0002974

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3446	0.276	-4.879	0.000	-1.885	-0.804
high_income	0.1050	0.208	0.505	0.614	-0.303	0.513
rel_apply	-1.8925	0.481	-3.937	0.000	-2.835	-0.950
rel_attend	0.8070	0.254	3.172	0.002	0.308	1.306

Socio-economic, application-related, and standardized test-related variables

Dep. Variable:	is_ivy	No. Observations:	1556
Model:	Logit	Df Residuals:	1551
Method:	MLE	Df Model:	4
Date:	Thu, 27 Jun 2024	Pseudo R-squ.:	0.02113
Time:	04:09:47	Log-Likelihood:	-437.67
converged:	True	LL-Null:	-447.12
Covariance Type:	nonrobust	LLR p-value:	0.0008238

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3221	0.288	-4.584	0.000	-1.887	-0.757
high_income	0.1250	0.221	0.566	0.571	-0.308	0.558
rel_apply	-1.8041	0.579	-3.118	0.002	-2.938	-0.670
rel_attend	0.7837	0.268	2.924	0.003	0.258	1.309
rel_apply_sat	-0.0964	0.357	-0.270	0.787	-0.796	0.603

Full geographic and weighted attendance variables — Logit Re

Dep. Variable:	is_ivy	No. Observations:	1556
Model:	Logit	Df Residuals:	1549
Method:	MLE	Df Model:	6
Date:	Thu, 27 Jun 2024	Pseudo R-squ.:	0.02667
Time:	04:09:47	Log-Likelihood:	-435.19
converged:	True	LL-Null:	-447.12
Covariance Type:	nonrobust	LLR p-value:	0.0005564

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4529	0.425	-3.421	0.001	-2.285	-0.621
par_income_bin	0.0033	0.006	0.520	0.603	-0.009	0.016
rel_apply	-1.6472	0.587	-2.804	0.005	-2.799	-0.496
rel_attend	0.5917	0.287	2.065	0.039	0.030	1.153
rel_apply_sat	-0.1727	0.372	-0.464	0.643	-0.902	0.557
attend_unwgt	90.9179	41.305	2.201	0.028	9.962	171.874
high_income	-0.1473	0.372	-0.396	0.692	-0.876	0.581

Modeling Q1

OLS Regression

OLS Regression Results

Dep. Variable:	is_ivy	R-squared:	0.014
Model:	OLS	Adj. R-squared:	0.010
Method:	Least Squares	F-statistic:	3.607
Date:	Thu, 27 Jun 2024	Prob (F-statistic):	0.00148
Time:	04:09:48	Log-Likelihood:	-197.94
No. Observations:	1556	AIC:	409.9
Df Residuals:	1549	BIC:	447.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1412	0.031	4.618	0.000	0.081	0.201
par_income_bin	0.0003	0.000	0.562	0.574	-0.001	0.001
rel_apply	-0.1020	0.037	-2.754	0.006	-0.175	-0.029
rel_attend	0.0410	0.021	1.975	0.048	0.000	0.082
rel_apply_sat	-0.0139	0.021	-0.675	0.500	-0.054	0.027
attend_unwgt	7.3656	3.503	2.103	0.036	0.495	14.236
high_income	-0.0162	0.030	-0.541	0.588	-0.075	0.042

Omnibus:	925.229	Durbin-Watson:	1.989
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5319.738
Skew:	2.951	Prob(JB):	0.00
Kurtosis:	9.872	Cond. No.	4.18e+04

ML Model Selection

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.902564	0.000000	0.000000	0.000000
Decision Tree	0.882051	0.375000	0.315789	0.342857
Random Forest	0.915385	0.727273	0.210526	0.326531
XGBoost	0.917949	0.650000	0.342105	0.448276
SVM	0.902564	0.000000	0.000000	0.000000
Ensemble	0.912821	0.833333	0.131579	0.227273

is_ivy	
0	0.913669
1	0.086331

Modeling Q2: OLS Regression Results pt 1

OLS Regression Results

Dep. Variable:	attend_unwgt	R-squared:	0.297
Model:	OLS	Adj. R-squared:	0.296
Method:	Least Squares	F-statistic:	218.5
Date:	Thu, 27 Jun 2024	Prob (F-statistic):	3.04e-118
Time:	23:31:25	Log-Likelihood:	7451.7
No. Observations:	1556	AIC:	-1.490e+04
Df Residuals:	1552	BIC:	-1.487e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0003	0.000	1.932	0.053	-4.06e-06	0.001
high_income	0.0015	0.000	12.590	0.000	0.001	0.002
rel_apply	-0.0009	0.000	-3.950	0.000	-0.001	-0.000
rel_attend	0.0015	0.000	10.346	0.000	0.001	0.002

Omnibus:	762.991	Durbin-Watson:	2.017
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5921.708
Skew:	2.151	Prob(JB):	0.00
Kurtosis:	11.534	Cond. No.	13.2

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Model - MSE: 4.0998462399994445e-06

OLS Model - R2 Score: 0.3568017472851164

Key socio-economic and application-related variables

Modeling Q2: OLS Regression Results pt 2

Dep. Variable:	attend_unwgt	R-squared:	0.306
Model:	OLS	Adj. R-squared:	0.304
Method:	Least Squares	F-statistic:	170.8
Date:	Thu, 27 Jun 2024	Prob (F-statistic):	2.95e-121
Time:	23:44:59	Log-Likelihood:	7461.5
No. Observations:	1556	AIC:	-1.491e+04
Df Residuals:	1551	BIC:	-1.489e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.534e-05	0.000	0.376	0.707	-0.000	0.000
high_income	0.0013	0.000	10.896	0.000	0.001	0.002
rel_apply	-0.0015	0.000	-5.568	0.000	-0.002	-0.001
rel_attend	0.0016	0.000	11.183	0.000	0.001	0.002
rel_apply_sat	0.0007	0.000	4.428	0.000	0.000	0.001

Omnibus:	782.776	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6514.662
Skew:	2.195	Prob(JB):	0.00
Kurtosis:	12.012	Cond. No.	16.0

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Model - MSE: 4.173439293479692e-06

OLS Model - R2 Score: 0.34525621102850423

Modeling Q2: OLS Regression Results pt 3

Dep. Variable:	attend_unwgt	R-squared:	0.407
Model:	OLS	Adj. R-squared:	0.405
Method:	Least Squares	F-statistic:	212.4
Date:	Thu, 27 Jun 2024	Prob (F-statistic):	1.02e-172
Time:	23:44:59	Log-Likelihood:	7583.5
No. Observations:	1556	AIC:	-1.516e+04
Df Residuals:	1550	BIC:	-1.512e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0005	0.000	-3.625	0.000	-0.001	-0.000
rel_apply	-0.0015	0.000	-6.134	0.000	-0.002	-0.001
rel_attend	0.0018	0.000	13.525	0.000	0.002	0.002
rel_apply_sat	0.0005	0.000	3.598	0.000	0.000	0.001
high_income	0.0012	0.000	11.033	0.000	0.001	0.001
public	0.0016	9.98e-05	16.224	0.000	0.001	0.002

Omnibus:	825.696	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	8157.820
Skew:	2.282	Prob(JB):	0.00
Kurtosis:	13.247	Cond. No.	16.2

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

OLS Model - MSE: 3.8045551312151395e-06

OLS Model - R2 Score: 0.40312805175948396

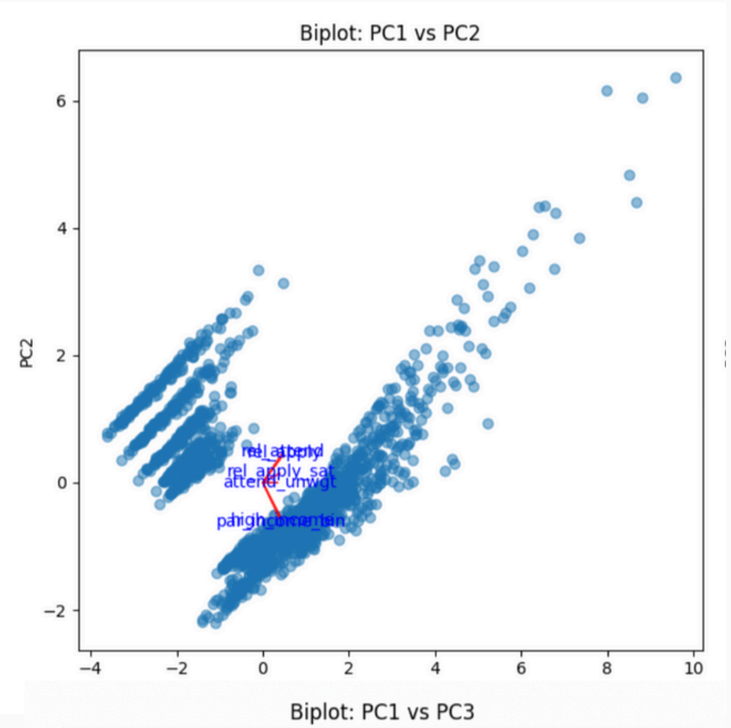
Modeling Q2

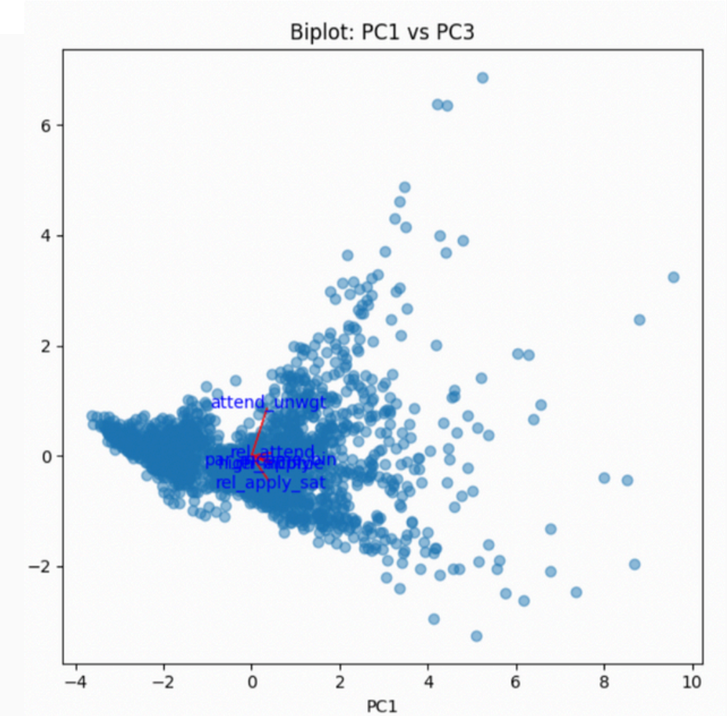
ML Model Selection

	Model	MSE	R2 Score
0	Linear Regression	0.000004	0.403128
1	Decision Tree	0.000007	-0.166296
2	Random Forest	0.000004	0.378225
3	XGBoost	0.000004	0.384629
4	SVM	0.000075	-10.744327

Improvements

- **Improve upon our data sources as we couldn’t find too many statistically significant variables in our prediction.**
- **Improve the way we’re segmenting variables (maybe use income as a continuous variable instead of categorical)**
- **Use class-imbalance models to improve beyond just the most common class**

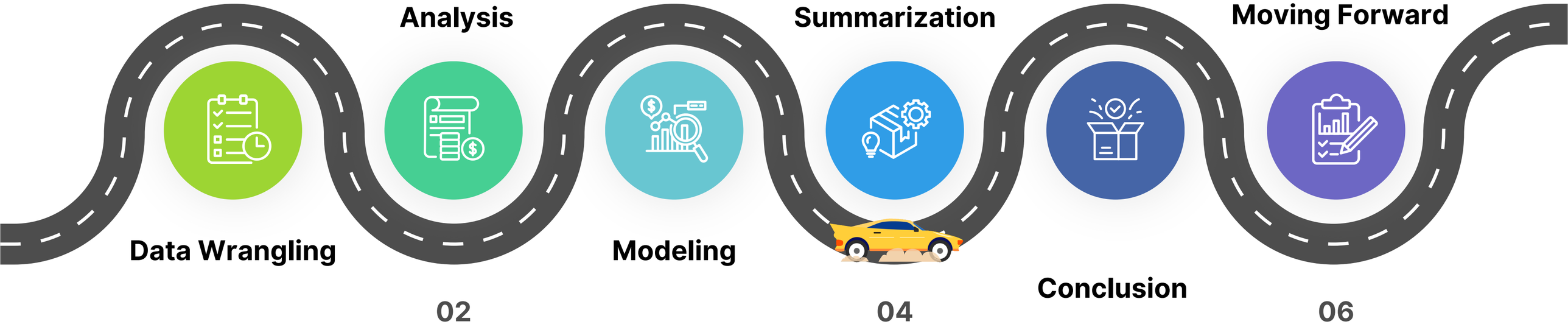




PCA

- **PC1 vs PC2:**
- **PC1 Influences: Dominated by attend_unwgt and rel_attend, indicating these attendance-related metrics significantly shape this principal component.**
- **PC2 Influences: Affected by high_income, rel_apply, and par_income_bin, highlighting the role of socioeconomic status and application behaviors in differentiating the data along this axis.**
- **PC1 vs PC3:**
- **PC3 Influences: Features contributions primarily from high_income and rel_apply_sat, suggesting these factors offer distinct, additional insights into data variance, especially related to economic status and standardized testing behaviors.**

Roadmap Research Process



Summary / Findings/ Conclusion

Key Insights from Modeling Analysis:

- High Income Influence:
  - No statistically significant impact on Ivy-Plus college admissions across all models.
  - Suggests admissions processes may not favor high-income applicants explicitly.
- SAT/ACT Submissions:
  - High statistically significant correlation with increased admission rates.
  - Indicates application rates that are weighted by SAT/ACT scores as decisive factors in attendance (because of it’s direct relationship)

Statistical and Model Observations for Q1

- Statistical Significance:
  - Few variables consistently showed statistically significant effects on admission chances.
- Model Convergence and Fit:
  - Models generally converged but had low explanatory power for the variability in admissions, suggesting potential limitations in the predictive accuracy of the models used.

Summary / Findings/ Conclusion cont.

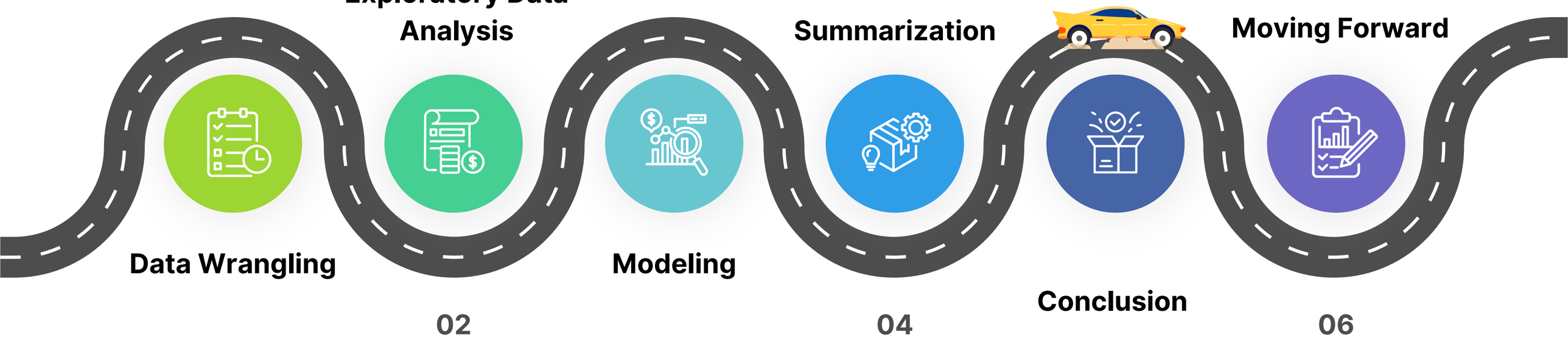
Implications for Stakeholders:

- Admissions Offices:
  - Should integrate holistic application reviews with an awareness of the significant impact of standardized test scores.
  - Consider balancing the weight given to SAT/ACT scores with qualitative aspects of student applications..
- Educators and Students:
  - Encouraged to maintain strong academic performance, particularly in standardized testing, while also developing well-rounded skills and experiences

Advice for Policy Makers and Educational Leaders:

- Increased Resource
  - Advised to ensure that test preparation resources are accessible to a broader range of students to support equity in college admissions.
- Program Development:
  - Encouraged to support educational initiatives that prepare students comprehensively, covering both standardized tests and holistic development.

Roadmap Research Process



Citations

- Opportunity Insights Data
- Images from Canva (Free Use)