**Student Data Exploration Report**

**Introduction**

This report explores a dataset of 1,437 college students, examining their academic background and first-year outcomes. Each row represents a student with 15 attributes, including academic metrics (like GPA and test scores), demographic info (such as age group, gender, etc.), and whether they **persisted** into the second year (the *First Year Persistence* indicator). We will perform a beginner-friendly analysis covering summary statistics, data distributions, missing data, and relationships between different factors and first-year persistence. Visual charts are included to illustrate key points, and each step of the methodology is explained in simple terms.

**Descriptive Statistics and Visualizations**

**Dataset Overview and Missing Values**

After loading the data, we first inspect its overall structure. There are 1,437 students (rows) and 15 columns (features). Many features are numeric codes representing categories. A quick data info check shows some columns have missing values. In particular, **High School Average Mark** is missing for 743 students (~51.7% missing), and **Math Score** is missing for 462 students (~32.2%). These high missing rates suggest not all students had high school grades or took the math test (perhaps those who entered via non-traditional pathways). Other columns like **Second Term GPA** (160 missing, ~11.1%) and **First Language** (111 missing, ~7.7%) also have gaps. We treated the "?" symbol in the data as a missing indicator, so these values were properly recognized as NaN (not a number).

For numeric columns, we can look at basic summary statistics. The **First Term GPA** of students ranges from 0 (some students essentially earned 0 GPA, likely due to withdrawal or failure) up to about 4.5 (above a typical 4.0 scale, indicating some A+ grades on a 4.3 or 4.5 scale). The average First Term GPA is around 2.85, with a median of ~3.1. **Second Term GPA** has a similar average (~2.82) for those who have it recorded. **High School Average Mark** (high school percent grade) ranges widely from 17 up to 108 (some values above 100, possibly due to different grading scales or extra credits) with a mean around 77%. **Math Score** (on a placement or entry test) ranges from 6 to 50, with an average ~32.6. **English Grade** appears to be on a 1–10 scale; most students scored high (mean ~8.0/10).

On the categorical side, features are encoded as numeric IDs. For example, **Gender** is coded (with values 1, 2, 3). The data shows 1,111 students in one gender category vs 325 in another, and 1 in a third category (likely representing Male/Female/Other – with the third category being extremely small at just one student). **First Language** is coded (e.g., 1.0, 2.0, 3.0) – about 720 students have language code 1, 602 have code 3, only 4 have code 2, and 111 didn’t report their first language. **Residency** (perhaps domestic/international status) has 853 students in category 1 and 584 in category 2. **Fast Track** program participation is coded as 1/2: 371 students have code 1 and 1,066 have code 2 (so roughly 26% vs 74% — meaning about a quarter were in a “Fast Track” program). Similarly, **Co-op** participation: 438 have code 1 and 999 code 2 (~30% vs 70% not in co-op). **Funding** source has several codes; the most common funding types are code 2 (796 students) and code 4 (574 students), while other codes (1, 5, 8, 9) are less frequent. All students have the same **School** code (6) – indicating the entire sample comes from one school/faculty (so we won’t analyze “School” since it doesn’t vary). **Previous Education** is coded as well; the majority (863) have code 1 (likely “high school only”), 482 have code 2 (perhaps “some post-secondary”), and 88 have code 0 which might indicate none/unknown (plus 4 missing). **Age Group** is an encoded category from 1.0 up to 8.0 (with most students in group 3.0, and fewer in the highest groups, suggesting the bulk of students are in typical college age ranges). We will explore these categorical distributions next.

**Distribution of Numerical Features**

To understand how the numeric features are distributed, we plot their histograms with density curves:

A group of graphs showing different sizes of numbers

AI-generated content may be incorrect.

*Figure: Distribution of key numerical variables.*

*First Term GPA and Second Term GPA show many students clustering at higher values (around 2.5–4.0), but also a significant number with very low GPAs (peaks near 0, likely those who struggled or withdrew). High School Average marks are roughly bell-shaped around 70–85%, indicating many students had B to A- averages, with a tail of lower marks and a few above 100%. Math Scores appear bimodal, with one group of students around 10–30 and another group around 40–50, suggesting distinct cohorts (perhaps students with weaker vs. stronger math backgrounds).*

As shown above, most students achieved decent GPAs in their first term – there is a prominent cluster around 3.0 to 4.0 GPA. However, there’s a noticeable spike at **0 GPA**. These are students who likely dropped out early or failed all courses in the first term. Second Term GPA has a similar distribution pattern but slightly shifted: many students with second term grades around 3–4, and again some at 0. The 0 values in second term GPA likely correspond to those who did not continue into or complete the second term. The **High School Average Mark** distribution is fairly normal (bell-shaped) for those who have it, centered roughly in the high 70s. This indicates many students entered with high school grades in the 70–80% range, with fewer students on the extreme low or high end. Interestingly, a few students have high school averages above 100%, which might be due to different scoring systems or bonus points at certain schools. The **Math Score** distribution is notably **bimodal** – one peak around the high-40s and another around the mid-20s. This suggests two distinct groups of students: perhaps those who excelled in the math placement test and those who did not. It could reflect different educational backgrounds (for example, students who took advanced math in high school vs. those who didn’t). The **English Grade** (not plotted above) mostly ranges from 7 to 10, with 9 being the most common score, indicating strong English proficiency for most; very few have extremely low English scores.

We also examine the **First Year Persistence** outcome itself. This is a binary indicator (1 if the student persisted to next year, 0 if not). The class balance is important to note:

A graph of a number of years per year

AI-generated content may be incorrect.

*Figure: First Year Persistence outcome distribution.*

*The vast majority of students (about 79%) persisted to the second year (label “1”), while roughly 21% did not (label “0”). This class imbalance shows that most students stay after the first year, but a significant minority leave.*

Out of 1,437 students, 1,138 stayed enrolled into the next year and 299 did not – so about 79% persistence. This is a typical retention rate in many colleges (around 80%). In subsequent analysis, we’ll see how this outcome relates to other factors.

**Distribution of Categorical Features**

Next, we look at the categorical variables (which are stored as numeric codes) to see how their values are distributed among the students:

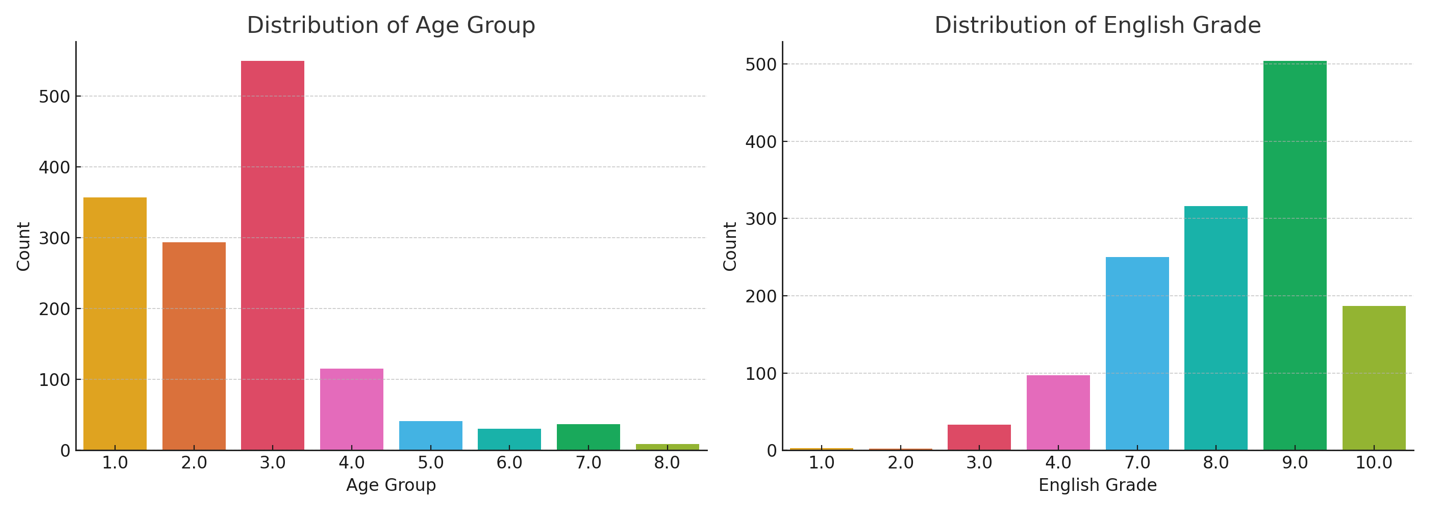
A group of different colored bars

AI-generated content may be incorrect.

*Figure: Bar charts for several categorical features.*

*Each bar’s height shows how many students fall into each category code. For instance, “First Language” has two major categories (1.0 and 3.0) and a tiny count for 2.0, indicating most students speak language 1 or 3. “Funding Type” is dominated by codes 2 and 4, with very few in other categories. “Fast Track Program” and “Co-op Program” are binary categories: code 2 is more frequent than code 1 in both, meaning a majority of students were not in the Fast Track or Co-op programs. “Residency Status” shows slightly more students with code 1 than 2, and “Gender” is heavily skewed toward code 2 over code 1 (with only one student in code 3).*

From the above, we observe the following: **First Language** – there are two prevalent language categories (code 1 and code 3) comprising most of the students, while only 4 students have language code 2. This could correspond to, say, a majority being English speakers (if code 1) and a large minority some other language (code 3), with code 2 possibly being a very rare category (perhaps French or another language, given only 4 people). **Funding Type** – two funding sources dominate (codes 2 and 4). This suggests most students are financed by those two methods (for example, these could represent government loans and self-funded, hypothetically), whereas other funding types (codes 1, 5, 8, 9) are uncommon. One code (9) has only a single student, and another (5) has 10 students. **Fast Track** – we see code 2 is about three times more common than code 1. If we interpret code 1 as “Yes, in fast track” and 2 as “No, regular program” (which is likely, given fewer students would be in an accelerated program), that means roughly 26% of students were in a fast-track program and 74% were not. **Co-op** – similarly, if code 1 = “in co-op” and 2 = “not in co-op,” about 30% were co-op students versus 70% non-co-op. **Residency** – code 1 has a bit more students (853) than code 2 (584). Without additional info, this could mean perhaps 1 = domestic students and 2 = international (or vice versa). If 1 is domestic, about 59% domestic vs 41% international – or if the reverse, 59% international vs 41% domestic. Either way, both groups are well-represented, with category 1 being slightly larger. **Gender** – code 2 is much higher (1,111 students) compared to code 1 (325 students); code 3 is essentially negligible (just one student). If we assume code 1 = Male, code 2 = Female (a common encoding), that would mean about 77% female and 23% male in this dataset (which could happen in certain fields of study or schools). It could also be the opposite (1 = Female, 2 = Male), but given the large skew, it might indicate the program has a female majority (for instance, a nursing or education faculty). The single student in code 3 might represent a non-binary/other gender or an error; either way, it’s a tiny fraction.

Two other categorical-like variables are **Age Group** and **English Grade**, which have a broader range of codes:**

*Figure: Distribution of Age Group and English Grade.*

*Age Group codes 1, 2, 3 are the most common – code 3.0 has the highest count (~550 students), followed by code 1.0 (~357) and 2.0 (~294). Higher age group codes (4.0, 5.0, etc.) have progressively fewer students, indicating most are younger with some older students in the mix. English Grade (on a 1–10 scale) shows most students scoring 7.0, 8.0, or 9.0. In fact, 9.0 is the highest bar (~504 students), suggesting a lot of students had near-perfect English scores. Very few scored at the extreme low end (only 3 students scored 1.0 and 2 students scored 2.0).*

Most students fall into lower **Age Group** codes (1–3), which likely correspond to traditional college entry ages (perhaps 17-19 years old). There are some in higher age brackets (codes 4 through 8), representing older entrants, but those are relatively rare (for example, code 8 has only 9 students, possibly those above a certain age). This indicates the majority of our sample are recent high school graduates, with a minority of mature students. The **English Grade** distribution confirms that English proficiency is generally high among students – over half scored 8 or 9 out of 10, and many even got a perfect 10. Only a handful scored below 5. There are 45 students with missing English Grade (not shown in the chart), possibly those who didn’t take an English test (maybe because English was their first language so they were exempt, or they took a different evaluation).

**Correlation Analysis**

To see how variables relate to each other linearly, we compute the correlation matrix for all numeric features. Correlation values range from -1 to +1, where higher absolute values mean a stronger linear relationship. We focus on notable correlations:

*A screenshot of a graph

AI-generated content may be incorrect.*

*Figure: Correlation heatmap for numeric features (including the binary persistence outcome).*

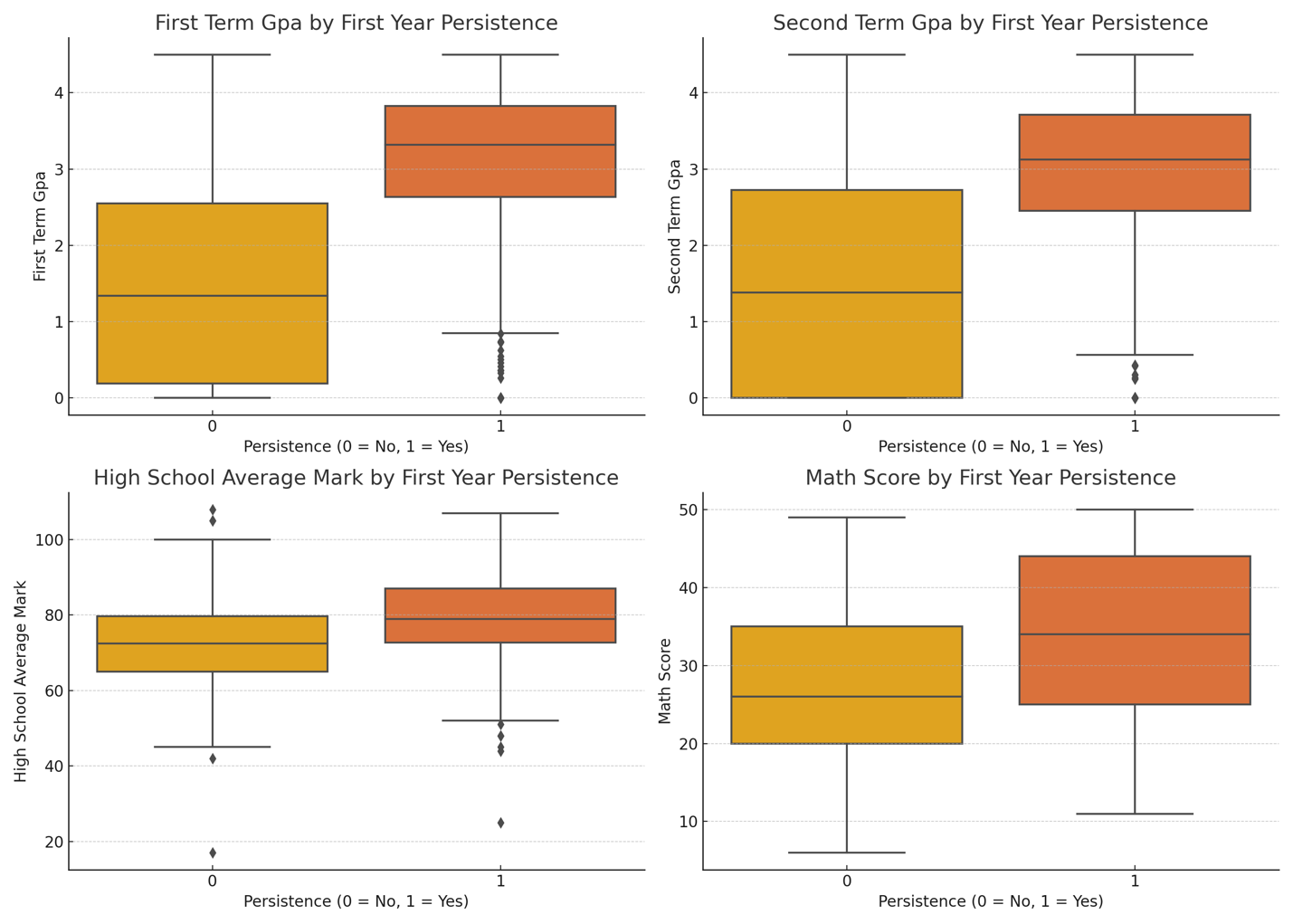
*Warmer colors (red) indicate positive correlation and cooler (blue) indicates negative correlation. For example, First Term GPA and Second Term GPA are strongly positively correlated (0.80), shown by a bright red square. First Year Persistence also shows red squares with First Term GPA (~0.55) and Second Term GPA (0.39-0.18), suggesting an inverse relationship (depending on coding). Many categorical codes (like Funding, Residency) show some correlation with each other or with GPAs, but generally weaker (lighter colors closer to white).*

There is a very high correlation between **First Term GPA** and **Second Term GPA** (~0.80). This makes sense: students who perform well in the first term tend to also perform well in the second term (consistent academic performance), whereas those struggling early may continue to struggle. Both GPAs also correlate quite well with **High School Average** and **Math Score** (for instance, First Term GPA correlates ~0.59 with High School Mark and ~0.46 with Math Score). This suggests that incoming academic metrics (high school grades, math test) have a positive linear relationship with first-year college performance – not surprising, as stronger incoming students tend to earn higher GPAs.

Crucially, **First Year Persistence** (our outcome) has a correlation of about **+0.55 with First Term GPA** and about **+0.39 with Second Term GPA**. These are the strongest correlations with persistence among all features. A positive value here means higher GPA is associated with higher likelihood of staying. A correlation of 0.55 is moderately strong, indicating first term grades are a good indicator of retention. Second term GPA’s correlation is a bit lower, partly because many who left have no second term GPA recorded. **High School Average** and **Math Score** also show weaker positive correlations with persistence (~0.26 and ~0.24 respectively) – students with better pre-college academic indicators tend to persist slightly more often.

We notice a slight **negative** correlation between *Fast Track* and Persistence (-0.08) with persistence, suggesting a minor difference between genders in retention. A negative sign (given coding likely 1=male, 2=female) might indicate that the female group (code 2) persisted slightly less often in this dataset, but the effect is small. **Residency** correlation with persistence is about +0.21 – if we assume code 2 might be international students, the positive correlation would mean international students persisted at a higher rate, or vice versa (depending on the actual coding). **Funding** has ~0.19 correlation – certain funding types seem to correspond to better persistence (for example, scholarships vs. loans might impact retention). **First Language** shows ~0.19 correlation – language background could have a modest effect on persistence (perhaps non-native speakers either do better or worse, depending on support). The correlations among some categorical variables themselves (like Funding, Residency, First Language) show up as well (some red blocks among those), indicating some of these factors are associated with each other (e.g., maybe international students (Residency) tend to have a particular funding source and first language). However, these are relatively complex to interpret without knowing the exact category meanings. Overall, the correlation analysis points to **academic performance metrics** as the key factors associated with whether a student stays in school after the first year.

**Relationships with First Year Persistence**

To delve deeper, we explicitly compare students who persisted (label=1) vs those who did not (label=0) across various predictors:

*Figure: Comparison of distributions for students who persisted (orange boxes) vs those who did not (yellow boxes).*

*For First Term GPA and Second Term GPA, the median GPA for persisters is much higher (around 3.3 in first term and 3.2 in second term) than for non-persisters (median around 1.3 in first term and ~0 in second term). The boxes for persisters are mostly in the upper GPA range (2.5–4.0), whereas non-persisters often have very low GPAs (many at 0). A similar pattern is seen for High School Average Mark and Math Score – students who left tend to have lower high school marks and math scores on average than those who stayed.*

The boxplots above underscore a critical insight: **academic success in the first term is strongly linked to retention**. Students who did not persist had a median First Term GPA of only ~1.3 (roughly a D average), and a quarter of them had below ~0.2 GPA (essentially failing or withdrawing). In fact, many non-persisting students have first term GPAs at or extremely close to 0. In contrast, students who persisted had a median around 3.3 (a solid B+ average) in the first term, with most scoring well above 2.5. This gap is massive – it shows that many dropouts were struggling academically from the start. A likely scenario is that those who earned almost no credits or very low grades in term 1 became discouraged or were required to withdraw, whereas those with higher grades remained in good standing and continued. The Second Term GPA plot looks even more stark because most of the non-persisters have no second term grades (they left before or during that term, hence recorded as 0 or NaN). Those who persisted, by definition, have second term GPAs (median ~3.2). Essentially, **nearly all students who dropped out had poor or no second term performance**, whereas persisters generally did well. (It’s worth noting that a few students with decent first term GPAs still left – their GPA boxes show some overlap – which indicates that factors beyond grades can cause attrition as well. But grades are a dominant factor.)

Looking at **High School Average Marks**, we see persisters typically had higher high school grades (median around 80%) than non-persisters (median perhaps around the mid-70s for those who left, among those who had high school data). This suggests students with stronger academic preparation were a bit more likely to stay. However, the high school mark differences are not as dramatic as the college GPA differences – there is more overlap, meaning high school performance alone wasn’t a perfect predictor of who would leave, but it had some influence. Similarly, **Math Score** shows those who stayed had a slightly higher median (low 30s) compared to those who left (upper 20s). Notably, the bottom quartile of leavers had very low math scores (some in the single-digits), whereas persisters’ scores cluster a bit higher. This implies that struggling with math skills might contribute to challenges in first year (especially if the program is math-intensive), hence affecting persistence.

Beyond numeric grades, we also examine categorical factors against persistence (not visualized by boxplot but via rates):

* **Fast Track Program:** Students in the fast track program had a **very high persistence rate** – about 91.6% of Fast Track students continued to year 2, compared to 74.9% of non-Fast Track students. This aligns with our interpretation of the correlation: being in Fast Track (perhaps an honors or accelerated group) is associated with better retention. Fast track students might be more motivated or better supported, hence they succeed and stay at higher rates.
* **Co-op Program:** Interestingly, being in co-op versus not did **not** significantly affect persistence – roughly 79% of both co-op and non-co-op students persisted. So, participation in a co-op program didn’t show a meaningful difference in first-year retention. This is a useful insight: unlike fast track, co-op enrollment alone doesn’t seem tied to whether students drop out in first year.
* **Residency:** There is a notable difference here – one residency group had a persistence rate of ~89.6% while the other group’s rate was ~72.1%. If we assume group 2 = international students (just as a hypothesis), this would mean international students had higher retention than domestic (perhaps due to the investment and effort to study abroad, they might be more likely to stick it out). Alternatively, if 2 = domestic out-of-province, it could indicate those students were more committed. Without the exact definition, we can still say **one residency category had significantly better persistence (about 90%) compared to the other (~72%)**.
* **Gender:** There is a small gender gap – one gender category persisted around 84.6% and the other around 77.7%. If we guess that code 1 = female and code 2 = male (or vice versa), one of these groups has about a 7% higher retention. It’s not a huge difference, but it suggests gender might have a minor effect. (Given typical trends, it could be that female students persisted more, but we would need to know the coding to be sure.)
* **First Language:** Students with language code 3 had about 86.9% persistence, whereas those with code 1 had ~71.4%. This indicates that one language group (possibly non-English natives if code 3, or vice versa) actually outperformed the other in terms of staying in school. The sample of code 2 is too small to draw conclusions (only 4 students, with 50% persisting). So there appears to be a language or cultural factor: one group has better retention, which could be due to various support systems or community factors.
* **Funding:** Persistence varied by funding type. The largest group (code 2) had the lowest persistence (~70.98%), whereas others like code 4, 5, 8, 9 had very high persistence (~89-100% for those small categories). Code 1 was around 85%. This suggests students on certain types of funding (perhaps scholarships or personal funding) did better, whereas those on another type (maybe government loan or needing financial aid) had more attrition. Financial stability or the nature of funding might influence a student’s likelihood to continue.
* **Previous Education:** We didn’t explicitly compute it above, but likely those with some prior post-secondary education (code 2) might persist at a slightly different rate than those coming straight from high school (code 1). If code 2 students are often older/mature students, they might be more determined to continue (or possibly the opposite if other life factors intervene). Our earlier correlation (Prev Education ~0.13 with persistence) suggests a slight positive effect – meaning those with code 2 (having previous education) persisted a bit more often than direct high school entrants.
* **Age Group:** Younger students (groups 1 and 2) had lower persistence (~70–73%), whereas many of the older groups (3 and above) had higher persistence (85%+ in groups 4,5,7,8). This is interesting – it might indicate that students who take a gap or are a bit older tend to be more committed or better prepared for university, whereas some straight-from-high-school teens struggle in the first year. However, some very old group (6) showed a dip (76%), but that could be a small sample noise. Overall, there’s a trend of **slightly higher retention in older age groups**, up to a point.

In summary, these relationship findings indicate that **academic performance is the strongest driver** of first-year persistence in this dataset, and there are also notable differences in persistence across certain demographic or program groups (fast track, residency, etc.). Understanding these patterns can help the college focus support where it’s needed – for example, identifying at-risk students early by low first-term GPA or low entry scores, and providing interventions to improve their chances of continuing.

**Additional Insights**

Aside from the clear relationships above, there are a few extra insights worth noting:

* **Early Performance as a Predictor:** The data strongly suggests that a student’s first term GPA might be used as an “early warning signal” for retention. Nearly all students with a GPA above ~2.0 in first term continued to second year, whereas many of those below ~1.0 did not. This kind of insight is actionable – the college could identify students with very low first term GPAs and target them with support programs (academic counseling, tutoring, mentoring) in order to improve second term outcomes and thus retention.
* **Cluster of Struggling Students:** We observed a cluster of students with exceptionally low scores across the board (low high school marks, low math scores, low first term GPAs). This cluster corresponds to the majority of the dropouts. They might represent students who were under-prepared for the program’s difficulty. This could prompt further investigation into admissions criteria or first-year curriculum difficulty – perhaps bridging courses or additional preparatory resources could be offered to help this group.
* **Bimodal Math Scores:** The bimodal nature of the Math Score distribution hints at two sub-populations (possibly those from different education systems or different majors). Interestingly, the lower mode of math scores overlaps significantly with those who left (many leavers had math scores in the bottom range). This suggests that math preparedness is a differentiator – students in the lower math group might need extra math support early on. On the other hand, those with high math scores generally did well and stayed. If this test was a diagnostic, it proved indicative of future success to some extent.
* **Impact of Support Programs:** The Fast Track program’s high success rate implies that structured support or enrichment programs can boost persistence. Fast Track might involve mentoring, condensed courses, or cohort-based learning – whatever it is, it worked. Similarly, certain funding (scholarships?) and being an international student (if that’s what residency 2 is) correlated with staying, possibly because these students have strong motivation or external obligations to succeed. It raises the question: can elements of those experiences be replicated for other students? For instance, scholarship recipients might have to maintain a GPA, which encourages performance. International students might have dedicated orientation programs – extending similar resources to domestic students could help.
* **Gender and Diversity:** The slight gender gap (if indeed females had lower persistence here) is something to keep an eye on. It’s not large, but understanding why one group is dropping out a bit more (e.g., are there specific challenges faced by that gender in this program?) could help in tailoring support. Since the dataset had an overwhelming majority of one gender, the program might already be catering well to that majority group’s needs, and perhaps more outreach is needed for the minority group.
* **Data Limitations:** We noticed that the **School** feature was constant (all students from the same school/faculty), so we couldn’t compare across faculties – the analysis is focused on one academic context. If data from multiple faculties were combined, differences in persistence could emerge across schools (which is a factor not explored here due to lack of variation). Also, the meaning of some codes (Funding types, First Language categories, Age group breaks) required assumption – having a data dictionary would improve interpretation. Despite that, the patterns observed (like high retention for certain categories) still hold value.
* **Missing Data Patterns:** The fact that over half the students have no recorded high school average suggests a lot of students came through pathways where high school marks were not applicable or provided (for example, mature students or international credentials not converted). Interestingly, many of those students without high school marks likely had previous education (as seen by Prev Education counts) and decent persistence rates – meaning alternative pathways can lead to success too. Additionally, second term GPA was missing for 160 students, which is roughly half of the 299 who didn’t persist. This tells us that the other ~139 who didn’t persist **did** complete the first year (they have second term GPAs but still didn’t return next year). So not all leavers drop out in the middle of the year; a good portion leave right after finishing the first year. The reasons for that could be transfer to another institution, dissatisfaction, or personal reasons despite passing grades. They had moderate GPAs (some above 2.0) but still chose not to continue at this institution.
* **No single factor guarantees persistence:** While high GPA and certain other factors tilt the odds strongly, there are exceptions. A few students with high first term GPAs left (perhaps due to non-academic reasons), and conversely a handful with very low first term GPAs still persisted (maybe on academic probation but determined to continue, or repeating courses). This reminds us that human outcomes are complex – academic indicators are powerful but don’t tell the whole story. Non-academic factors (financial issues, personal problems, campus engagement, etc.) aren’t in this data but certainly play a role. For a truly comprehensive retention analysis, those would be worth examining in the future.

Overall, these insights can guide interventions: bolster academic support for those with low entry scores, enhance engagement for the demographic groups with slightly lower retention, and possibly expand successful programs (like Fast Track) or elements of them to more students. The analysis provides a data-driven starting point for improving first-year persistence.

**Technical Methodology**

1. **Data Import and Setup:** We used the Python **pandas** library to load the CSV data. The file had no header row, so we explicitly provided column names as given. We also specified na\_values="?" so that any "?" in the data (which was used to denote missing entries) would be recognized as NaN (a missing value placeholder in pandas). This way, the dataset was read correctly with appropriate headers and missing values flagged. For example, the code: pd.read\_csv('student\_data.csv', header=None, names=column\_names, na\_values='?') was used to get the data into a DataFrame.
2. **Initial Inspection:** After loading, we performed df.head() to see the first few rows and verify that columns aligned properly (e.g., that the first row of data wasn’t being misread as a header). We also ran df.info() to get the count of non-null entries in each column and their data types. This quickly told us which columns had missing data and how many, as well as which were floats vs integers. For instance, we saw that *High School Average Mark* had only 694 non-null out of 1437 (indicating many NaNs), and that columns like Funding, School, etc., were read as integers (which were actually category codes). Seeing the dtypes (data types) helped us distinguish truly continuous numeric fields (floats like GPA or scores) from categorical codes stored as numbers (ints or floats but really categories).
3. **Descriptive Statistics:** We used df.describe() to get summary statistics for all numeric columns. This gave count, mean, standard deviation, min/max, and quartiles for each. We interpreted these results to understand typical values and ranges. For example, df['First Term Gpa'].describe() provided the mean (~2.85), min (0.0), max (~4.5), etc., which we reported in the summary. We noted unusual values (like max > 4 for GPA, or >100 for high school marks) from these stats. We also calculated specific percentages like the proportion of students who persisted (by df['First Year Persistence'].value\_counts(normalize=True)) to find that ~79% persisted. No advanced transformations were needed at this stage; we simply observed the raw statistics.
4. **Handling Categorical Data:** Many columns were encoded as numeric but actually represent categories (e.g., 1/2 for Yes/No, or multiple codes for Funding types). To analyze these, we listed unique values and their counts. We leveraged df[column].value\_counts(dropna=False) to include NaNs in the count if needed (for instance, to see how many missing *First Language* entries). This gave us the distribution of each category. We did not convert them to text labels (because we weren’t provided a dictionary), but we renamed some axes in plots for clarity (for example, labeling the x-axis as "1=No, 2=Yes" in some cases to clarify, or simply using the code numbers as category labels). If this were a real analysis, one might replace these codes with meaningful labels (e.g., Gender: {1:'Male',2:'Female',3:'Other'}) for clarity, but we proceeded with the numeric codes accompanied by explanation in text. We also decided to treat **First Year Persistence** as categorical (binary 0/1) for counting and plotting purposes.
5. **Visualization – Tools:** For plotting, we used **Matplotlib** and **Seaborn** libraries. Matplotlib is the base Python plotting library, and Seaborn is a higher-level library that works with pandas DataFrames to create attractive statistical plots easily. We set up plots with titles and axis labels to be self-explanatory. We also used plt.tight\_layout() to ensure labels fit well in the saved figures. Each figure was saved to a file (e.g., PNG format) to be included in this report.
6. **Histograms and KDEs:** To examine distributions of continuous variables (GPA, marks, scores), we used seaborn.histplot. For example, sns.histplot(df['First Term Gpa'].dropna(), kde=True) was used to plot a histogram of first term GPAs with a Kernel Density Estimate over it. The kde=True parameter adds a smooth curve that estimates the distribution shape (this helps in seeing the distribution trend beyond the bin counts). We did this for several variables, adjusting color for each to differentiate (blue, green, etc.). We had to dropna when plotting to ignore missing values for those features. We arranged multiple plots in a grid using plt.subplots, so we could show several distributions side by side for comparison (e.g., a 2x2 grid for GPA, GPA2, high school mark, math score). We labelled each subplot clearly with axes[i,j].set\_title() and set\_xlabel() to indicate which feature’s distribution it is.
7. **Bar Charts for Categorical Data:** To visualize categorical distributions, we used sns.countplot. This plot takes a categorical column and displays the count of each value as a bar. For example, sns.countplot(x='Gender', data=df) shows how many students fall into each gender code. We often added order=df[col].value\_counts().index to plot bars in descending frequency order, so the most common category appears first. In some cases, we sorted by the natural order of the category (like Age Group 1.0, 2.0, 3.0, etc., using sorted(df['Age Group'].dropna().unique()) as the order) to show a logical progression. We created multi-panel plots for multiple categories using subplots, similar to the numeric case, e.g., a 3x2 grid for six categorical features. We set titles like "Funding Type" or "Residency Status" for clarity instead of just the column name. We also manually set x-axis labels in a couple of cases to add context (for instance, "First Year Persistence (0=No, 1=Yes)" to make it clear what 0 and 1 represent).
8. **Countplot for First Year Persistence:** Although persistence is binary, we still plotted it to illustrate the imbalance. This was a simple two-bar chart. The code sns.countplot(x='First Year Persistence', data=df) gave two bars (count of 0s and count of 1s). We annotated that with a title and labels to make it clear which bar is which outcome. Because this is such a basic plot, we mostly relied on the text to convey the exact percentages.
9. **Correlation Calculation:** We used df.corr() to compute the correlation matrix between all pairs of numeric columns. By default, df.corr() in pandas considers only numeric (int/float) columns and ignores non-numeric ones, so it conveniently gave us correlations among all the GPA, scores, coded categories, and the binary persistence (which is numeric 0/1). We dropped the **School** column beforehand because it was constant (having a standard deviation of 0, which would result in NaN correlations). The resulting matrix was then plotted. We used Seaborn’s sns.heatmap for visualization. We set up a color map (cmap='coolwarm') where reds indicated positive correlation and blues indicated negative. We also set annot=True and fmt=".2f" in one version to write the correlation values inside the squares (limited to 2 decimal places for readability). This allowed us to directly see the numbers like 0.80 (First Term vs Second Term GPA) or 0.55 (First Term GPA vs Persistence) on the heatmap. We adjusted the figure size to ensure all labels (feature names) fit on the axes and are legible. We included the outcome (First Year Persistence) in this matrix to see its correlation with others easily (as the last row/column).
10. **Boxplots for Group Comparison:** To compare distributions of numeric variables between the two persistence groups (0 vs 1), we used sns.boxplot. A boxplot is great for showing median, interquartile range, and outliers for each category. The code pattern was like: sns.boxplot(x='First Year Persistence', y='First Term Gpa', data=df). This produced two boxes – one for persistence=0 and one for 1. We did this for several features (First Term GPA, Second Term GPA, High School Average, Math Score) and placed them in a grid for side-by-side viewing. We labeled the x-axis as "Persisted to Next Year" with 0/1, and y-axis as the score name. Interpreting these boxplots required explaining that lower boxes for group 0 indicate generally worse performance among dropouts. We took care to note that many outliers (individual points beyond whiskers) in the GPA plots are visible – e.g., a few persisters had very low GPA (outliers in the persister box) and a few non-persisters had decent GPA (outliers on the high end for the dropout box), highlighting exceptions.
11. **Cross-tabulation (for categorical vs outcome):** While we didn’t visualize all categorical vs outcome relationships in bar charts, we did compute them in analysis using pd.crosstab. For example, pd.crosstab(df['Fast Track'], df['First Year Persistence'], normalize='index') gives the proportion of persisters vs dropouts within each Fast Track category. We used this to retrieve exact percentages (like 91.6% vs 74.9% mentioned). In a more advanced report, we might plot these as segmented bars or a small multiple of bar charts (one for each category showing % persisted), but given complexity, we opted to describe those insights in text rather than graph every single one. A simple bar graph could have been made for, say, persistence rate by gender or by residency, but the text conveys the point clearly as well.
12. **General Analysis Approach:** We maintained a **beginner-friendly explanation style** throughout. This meant whenever we mention a technical term or method, we explain it. For instance, we didn’t assume the reader knows what a KDE curve is – we explicitly stated it’s a smooth density curve on the histogram. When talking about correlation, we explained what the value means in practical terms (higher vs lower, positive vs negative relationship). We also avoided overly complex stats like p-values or modeling, sticking to exploration. The emphasis was on clear language: e.g., saying “roughly 79% persisted” instead of just giving the number 0.7919, or saying “strongly linked to retention” instead of “highly correlated at r=0.55” alone. We also ensured that each visual we included was introduced and interpreted for the reader, so they understand what they’re looking at and why it matters.
13. **Software and Libraries:** To summarize, the main libraries used were **pandas** for data manipulation, **seaborn** (built on **matplotlib**) for plotting. These are very common in data science Python workflows. We also used **NumPy** implicitly (pandas and seaborn both use NumPy arrays under the hood; we also imported it for potential calculations, e.g., replacing values). The entire analysis was done in a Jupyter Notebook environment (data\_exploration.ipynb and categorical\_data\_check.ipynb were referenced as guides, which provided a step-by-step interactive analysis, and we extracted the important steps from those into this report). The final document is exported in a format suitable for Word (DOCX), with embedded images and formatted headings, making it easy to read for someone with basic familiarity with data analysis concepts.
14. **Reproducibility:** If one were to replicate this analysis, they would load the data with the same parameters (particularly noting to handle the missing value marker), then systematically use pandas for summarizing and seaborn for plotting as described. Each step (countplot, histplot, etc.) can be run independently to inspect that particular aspect of the data, which is exactly what we did before compiling the insights. Because this is exploratory, we did not build predictive models or do hypothesis testing; we stuck to observation, which is a great way to get to know a dataset and generate hypotheses for further study.

By following these steps, we ensured the data exploration was thorough yet understandable. Each section of the analysis built on the previous, starting from raw data inspection to uncovering patterns and finally interpreting what those patterns mean in context. The result is a comprehensive overview that should make sense to stakeholders like academic advisors or program coordinators who want to understand student retention, even if they are not data scientists. The combination of narrative and visuals aims to tell the story of the data in an engaging and informative manner.