Predictive Modeling and Spatial Analysis Report

Author: Kenier Ramirez

Date: November 3, 2025

Research Question

The research question guiding this study is: **Can a logistic regression model be constructed on the San Diego County Business Listings dataset to estimate the likelihood of business survival?** This question is justified by the increasing need for data-driven tools that support entrepreneurs and investors in understanding which factors influence business longevity. Given that business survival is shaped by demographic, geographic, and economic characteristics, developing a statistical model that captures these relationships can provide meaningful insights and practical value.

The context for this research is grounded in prior studies that highlight the importance of spatial and structural factors in firm survival. For example, Reyes and Suárez (2022) demonstrated that a business's survival probability is strongly associated with the characteristics of the road network on which it is located, suggesting that locational context and environmental features play a significant role in long-term operational viability. Building on this foundation, the present study employs logistic regression on a comprehensive dataset of San Diego County businesses to identify significant predictors of account status (active versus inactive) and to assess whether these predictors can be used to construct an accurate and interpretable model.

The hypothesis for this analysis is twofold.

- **Null Hypothesis (H_0):** A predictive logistic regression model **cannot** be constructed from the research dataset.

- **Alternative Hypothesis (H_1):** A predictive logistic regression model **can** be constructed from the research dataset at a model accuracy of **70% or higher**.

To test this hypothesis, the study utilizes a publicly available dataset containing 146,481 business records (Ramirez, 2025). Each record includes variables such as account creation date, ownership type, NAICS sector and code, industry description, geographic identifiers (city, ZIP code, latitude, longitude), population estimates, and the target variable *account_status*. These features represent relevant economic, structural, and spatial characteristics that may influence business survival. By analyzing these variables through logistic regression, the study seeks to determine both the feasibility and the predictive strength of a model designed to estimate business survival outcomes.

Data Collection

The data used in this study were collected from publicly available CSV files. Data was integrated using the following three primary public data sources: the City of San Diego's Active and Inactive Business Listings datasets (City of San Diego, n.d.) and the 2020 Census ZIP Code–level population dataset compiled by the San Diego Association of Governments (SANDAG, n.d.). The final dataset contains **146,481 registered businesses** established between July 1, 1974, and November 3, 2025. Each record includes key variables such as account creation date, ownership type, NAICS sector and code, city, ZIP code, latitude, longitude, population, and the response variable *account_status*, which indicates whether a business is active or inactive. Several derived variables were also generated to support the analysis, including business age, business life-cycle groupings, local business density, and active/closed rates by ZIP code, NAICS sector, or city. These derived measures align with accepted practices in multivariate analysis and predictive modeling.

An important **advantage** of this data-gathering methodology is that the dataset draws from reputable, publicly maintained sources, ensuring transparency, accessibility, and consistent documentation. Because it integrates three independently curated datasets, it provides both business-level detail and contextual demographic information, supporting a more comprehensive modeling framework. Moreover, using official municipal and Census data ensures high validity, broad geographic coverage, and reliable identifiers (ZIP codes, NAICS codes). This allows the study to investigate economic and geographic predictors simultaneously, increasing the robustness and interpretability of the logistic regression model.

However, the methodology also presents a significant **disadvantage**. The dataset reflects only *registered* businesses, excluding unregistered or informal establishments. This limitation may result in an underrepresentation of true business density in certain ZIP codes, potentially reducing the generalizability of findings to the broader business ecosystem. Additionally, the dataset includes some imprecision in latitude and longitude values and relies on ZIP Code–level population counts from the 2020 Census (SANDAG, n.d.), which do not capture population changes occurring after 2020. In addition, business registries often contain inconsistent address formatting, missing values, and category inconsistencies**.** These limitations represent constraints on both spatial accuracy and temporal relevance.

To address **the challenges** encountered during the data collection and preparation process, several steps were taken to mitigate them. Python was used extensively for data cleaning, transformation, and validation. Missing values were imputed where appropriate, categorical variables were converted into binary dummy variables, and duplicate or invalid records were removed. Outliers in *date_account_creation* were retained as delimitations of the study to preserve potentially meaningful cases of long-standing businesses. Additionally, sectors 22 (Utilities) and 92 (Public Administration) were excluded to ensure that the analysis remained focused on private-sector business activity. These measures ensured that, despite limitations inherent in the source data, the final dataset was sufficiently clean, consistent, and analytically ready for logistic regression modeling (Ramirez, 2025).

Challenges & How They Were Overcome

Challenge	Strategy Used
Missing values & incomplete entries	Explicit NA filtering, dropping nulls where appropriate, and imputations were avoided to preserve data integrity.
Inconsistent city names	Manual standardization + checking for out-of-county records.
NAICS sector formatting	Creation of standardized “naics_sector_standard” categories.
Merging shapefiles for filtering	Used geopandas to spatially filter businesses with valid lat/long coordinates
Duplicate entries	Removed via drop_duplicates()

Data Extraction and Preparation

1. Data Extraction

The data used in this study were collected from publicly available CSV files (Ramirez, 2025), which consolidate multiple authoritative public data sources, including the City of San Diego Active and Inactive Business Listings datasets (City of San Diego, n.d.) and the 2020 ZIP Code–level population dataset from the San Diego Association of Governments (SANDAG, n.d.).

1.1 Load Active & Inactive Business Data:

```
# ======================================
# Load active and inactive datasets
# ======================================
active_df = pd.read_csv(
    "sd_businesses_active_datasd.csv",
    low_memory=False,
    parse_dates=['date_account_creation', 'date_cert_expiration', 'date_cert_effective', 'date_business_start']
)
inactive_df = pd.read_csv(
    "sd_businesses_inactive_2015tocurr_datasd.csv",
    low_memory=False,
    parse_dates=['date_account_creation', 'date_cert_expiration', 'date_cert_effective', 'date_business_start']
)
```

Merged datasets by stacking them, then removing duplicates.

```
# ======================================
# Combine active and inactive datasets
# ======================================
business_df = pd.concat([active_df, inactive_df], ignore_index=True)
```

```
# Business df info
business_df.head()
```

	account_key	account_status	date_account_creation	date_cert_expiration	date_cert_effective	business_owner_name	ownership_type
0	1974000024	ACTIVE	1974-07-01 12:00:00	2026-06-30 12:00:00	2025-07-01 12:00:00	PARRON HALL CORP	CORP
1	1974000035	ACTIVE	1974-07-01 12:00:00	2026-06-30 12:00:00	2025-07-01 12:00:00	UNIV MECHANICAL & ENGINEERING CONTRACTORS INC	CORP
2	1974000039	ACTIVE	1974-07-01 12:00:00	2026-06-30 00:00:00	2025-07-01 00:00:00	ADMIRAL EXCHANGE CO INC	CORP
3	1974000053	ACTIVE	1974-07-01 12:00:00	2026-06-30 00:00:00	2025-07-01 00:00:00	R W SMITH & CO INC	CORP
4	1974000080	ACTIVE	1974-07-01 12:00:00	2026-06-30 12:00:00	2025-07-01 12:00:00	ALDERWOODS GROUP INC	CORP

5 rows × 27 columns

1.2 Remove Pending Businesses

Filtered account_status = "PENDING" to keep only ACTIVE or CLOSED businesses.

account_status	count
CANCELLED	96960
ACTIVE	57919
PENDING	3493

dtype: int64

- **Drop pending businesses**

```
# Drop Pending business status
business_df = business_df[business_df['account_status'].str.upper() != 'PENDING']

# Check Counts
business_df['account_status'].value_counts()
```

account_status	count
CANCELLED	96960
ACTIVE	57919

1.3 Remove Duplicates, Drop Irrelevant Columns & Check Missing Data

```
# ====================================
# Remove Duplicates
# ====================================

# Numbers of rows
print("Original Number of Rows")
print(len(business_df))

# Remove duplicate records based on account_key
business_df= business_df.drop_duplicates(subset=['account_key']).reset_index(drop=True)
print("Number of Rows after Removal")
print(len(business_df))

# ====================================
# Drop irrelevant variables
# ====================================

drop_cols = ['account_key','business_owner_name', 'dba_name','address_no','address_pd','address_road',
             'address_sfx','address_no_fraction','address_suite', 'address_pmb_box', 'address_po_box', 'bid', 'council_dist
             'date_cert_expiration', 'date_cert_effective']
business_df.drop(columns=[c for c in drop_cols if c in business_df.columns], inplace=True)

# Account status count
business_df['account_status'].value_counts()
```

```
Original Number of Rows
158372
Number of Rows after Removal
158372
```

Missing-value summary

```
# Check for missing values
missing_business_df = business_df.isnull().sum().sort_values(ascending=False)
missing_percent_business_df = (missing_business_df / len(business_df) * 100).round(2)
pd.DataFrame({'Missing': missing_business_df, 'Percent': missing_percent_business_df})
```

	Missing	Percent
lng	3400	2.2
lat	3400	2.2
address_city	1	0.0
account_status	0	0.0
date_business_start	0	0.0
ownership_type	0	0.0
date_account_creation	0	0.0
naics_sector	0	0.0
naics_description	0	0.0
naics_code	0	0.0
address_zip	0	0.0
address_state	0	0.0

1.4 Geospatial Filtering with Shapefiles

Spatial join to limit businesses to the boundaries of San Diego County. Because the dataset contains latitude and longitude values that occasionally include transcription errors (e.g., missing minus signs, invalid coordinates), a bounding-box filter was applied prior to spatial processing.

```
# ======================================
# Drop missing or invalid coordinates
# ======================================
business_df = business_df.dropna(subset=['lat', 'lng'])

# Keep only realistic SD County bounding coordinates
business_df = business_df[
    (business_df['lat'].between(32.5, 33.5)) &
    (business_df['lng'].between(-117.5, -116.0))
]

# Convert to GeoDataFrame with point geometry
gdf = gpd.GeoDataFrame(
    business_df,
    geometry=gpd.points_from_xy(business_df['lng'], business_df['lat']),
    crs="EPSG:4326"
)
```

A preliminary geographic bounding-box filter was used to remove invalid or implausible coordinate values. San Diego County lies approximately within:

- **Latitude:** 32.5° to 33.5°
- **Longitude:** −117.5° to −116.0°

Points falling outside these ranges indicate data-entry errors such as:

- Latitudes of 0, 20, or 90
- Longitudes of +117 instead of −117
- Reversed or swapped degree values
- Coordinates mistakenly placed in Mexico or out of state

These erroneous points cannot be assigned to any Census county or ZIP polygon. These rows would be dropped by spatial joins anyway; the bounding-box filter removes them earlier, improving efficiency and preventing geometry errors.

The bounding-box filter **does not** assign ZIP codes or determine geography. Actual location assignment occurs through two authoritative TIGER/Line shapefiles:

- **Layer 1:** San Diego County boundary
- **Layer 2:** 2020 ZIP Code Tabulation Areas (ZCTAs)

```
# ======================================
# LAYER 1 — Filter to San Diego County
# ======================================
county_shp = gpd.read_file("zip://cb_2022_us_county_500k.zip")

sd_county = county_shp[
    (county_shp['STATEFP'] == '06') &
    (county_shp['COUNTYFP'] == '073')
]

gdf_sd = gpd.sjoin(gdf, sd_county, predicate='within', how='inner')

# Drop join artifacts so they don't conflict with layer 2
gdf_sd = gdf_sd.drop(columns=[col for col in gdf_sd.columns if col.startswith("index_")], errors='ignore')

# ======================================
# LAYER 2 — Assign ZCTA (ZIP Code)
# ======================================
zcta = gpd.read_file("zip://cb_2020_us_zcta520_500k.zip")

zcta = zcta.to_crs(gdf_sd.crs)

gdf_zcta = gpd.sjoin(
    gdf_sd,
    zcta[['ZCTA5CE20', 'geometry']],
    predicate='within',
    how='left'
)
```

Spatial join operations produced columns such as index_left and index_right.

```
# Drop any new join artifacts
gdf_zcta = gdf_zcta.drop(columns=[col for col in gdf_zcta.columns if col.startswith("index_")], errors='ignore')

# Rename ZIP column to zip5
gdf_zcta = gdf_zcta.rename(columns={'ZCTA5CE20': 'zip5'})

# Ensure zip5 is a 5-character string
gdf_zcta['zip5'] = gdf_zcta['zip5'].astype(str).str.zfill(5).str.strip()

# KEEP ONLY BUSINESSES THAT HAVE A ZIP (TO merge with Census ZIP population )
gdf_zcta = gdf_zcta.dropna(subset=['zip5'])

# ======================================
# Final dataframe
# ======================================
df = pd.DataFrame(gdf_zcta.drop(columns='geometry'))
```

The index columns created by the spatial join (index_left and index_right) were dropped because they are only internal references to polygon indices and are not relevant for assigning county membership or ZCTA codes.

Geospatial filtering ensured that non-San Diego businesses (resulting from malformed addresses or mis-entered coordinates) were excluded.

Advantage: High precision in geographic assignment.
Disadvantage: Records with missing or faulty coordinates had to be removed, even if they were valid.

1.5 Merge with Census Data. Merged on the ZIP code

```
# ======================================
# Load 2020 Census population
# ======================================
census_df = pd.read_csv(
    "2020_Census_Population_by_ZIP_Code_20251109.csv",
    low_memory=False, thousands=','   # <-- tells pandas to remove commas
    )

# ======================================
# Aggregate population by ZIP
# ======================================
# Sum all population entries for the same ZIP
census_agg = census_df.groupby('zip', as_index=False)['population'].sum()

# Convert ZIP to string to match DF dataset
census_agg['zip'] = census_agg['zip'].astype(str).str.zfill(5)
df['zip5'] = df['zip5'].astype(str).str.zfill(5)

# ======================================
# Merge with DF dataset
# ======================================
# 'zip5' in DF matches 'zip' in Census
df = df.merge(census_agg, left_on='zip5', right_on='zip', how='left')

# Drop duplicate 'zip' column after merge
df = df.drop(columns=['zip'])
```

```
Merged DF shape: (142472, 13)
```

Removal of Unmatchable ZIP Codes: A small number of businesses could not be matched to any 2020 ZCTA.

```
# Check missing ZIPs
missing_pop_zips = df[df['population'].isna()]['zip5'].unique()
print("ZIPs with missing population:", missing_pop_zips)
```

```
ZIPs with missing population: ['00nan' '92092' '92132' '92147']
```

```
# ======================================
# Handle ZIPs with missing population
# ======================================

# Define ZIPs
invalid_zip = '00nan'
real_missing_pop_zips = ['92092', '92132', '92147']

# Check how many rows would be affected
rows_invalid = df[df['zip5'] == invalid_zip].shape[0]
rows_missing_pop = df[df['zip5'].isin(real_missing_pop_zips)].shape[0]
print(f"Rows with invalid ZIP ('{invalid_zip}'): {rows_invalid}")
print(f"Rows with missing population (real ZIPs): {rows_missing_pop}")

# Drop invalid ZIP
df = df[df['zip5'] != invalid_zip].reset_index(drop=True)

# Assign zero population to real missing ZIPs
df.loc[df['zip5'].isin(real_missing_pop_zips), 'population'] = 0
```

A total of 33 out of 142,472 records (0.02%) had ZIP codes that did not correspond to any valid ZCTA. These rows were removed to ensure accurate population assignment. All removed cases were logged for reproducibility.

Handling Missing Data:

- Missing or corrupted population values were imputed where possible
- Rows with irreconcilable missing ZIP or coordinate data were removed.

```
# ======================================
# Explanation
# ======================================
explanation = f"""
The following ZIP codes were handled due to missing population:

- Dropped invalid ZIP: {invalid_zip}
- Assigned population = 0 to real San Diego County ZIPs: {', '.join(real_missing_pop_zips)}

This ensures that only invalid ZIPs are removed while keeping legitimate ZIPs for analysis.
"""
print(explanation)
```

```
Rows with invalid ZIP ('00nan'): 12
Rows with missing population (real ZIPs): 21

The following ZIP codes were handled due to missing population:

- Dropped invalid ZIP: 00nan
- Assigned population = 0 to real San Diego County ZIPs: 92092, 92132, 92147

This ensures that only invalid ZIPs are removed while keeping legitimate ZIPs for analysis.
```

1.6 Removal of Irrelevant Sectors and Columns: Based on research delimitations

- NAICS sectors **22 (Utilities)** and **92 (Public Administration)** were removed because they are not private-sector businesses.

```
# ======================================
# Sector 22: 'Utilities' &  92: 'Public Administration'
# Exclude sectors that are not business-related (update df in place)
# ======================================

df = df[~df["naics_sector"].isin([22, 92])]
```

- The variable **naics_description** was excluded since the NAICS code already encodes the same information. Additionally, the NAICS code was standardized for improved readability.

```
# ======================================
# Define mapping of 2-digit NAICS codes to human-readable representation in the economic sectors
# ======================================

naics_sector_map = {
    11: 'Agriculture, Forestry, Fishing and Hunting',
    21: 'Mining, Quarrying, and Oil and Gas Extraction',
    23: 'Construction',
    31: 'Manufacturing',
    32: 'Manufacturing',
    33: 'Manufacturing',
    42: 'Wholesale Trade',
    44: 'Retail Trade',
    45: 'Retail Trade',
    48: 'Transportation and Warehousing',
    49: 'Transportation and Warehousing',
    51: 'Information',
    52: 'Finance and Insurance',
    53: 'Real Estate and Rental and Leasing',
    54: 'Professional, Scientific, and Technical Services',
    55: 'Management of Companies and Enterprises',
    56: 'Administrative and Support and Waste Management and Remediation Services',
    61: 'Educational Services',
    62: 'Health Care and Social Assistance',
    71: 'Arts, Entertainment, and Recreation',
    72: 'Accommodation and Food Services',
    81: 'Other Services (except Public Administration)',
    }

# Function to convert full NAICS code to sector
def naics_to_sector_label(code):
    if pd.isna(code):
        return 'Unknown'

    # Convert to int (in case it is string)
    code = int(str(code)[:2])
    return naics_sector_map.get(code, 'Other')

# Apply mapping to dataset
df['naics_sector_standard'] = df['naics_code'].apply(naics_to_sector_label)

# Convert to uppercase
df['naics_sector_standard'] = df['naics_sector_standard'].str.upper()

# Check results
print(df[['naics_code','naics_sector_standard']].drop_duplicates().head(10))
df['naics_sector_standard'].value_counts()
```

```
# Check missing values specifically for NAICS columns
print("\nMissing values for NAICS columns:")
print(df[['naics_code', 'naics_sector_standard']].isna().sum())

# Which rows are missing NAICS codes
missing_naics = df[df['naics_code'].isna()]
print("\nRows with missing NAICS codes:")
display(missing_naics)
```

```
Missing values for NAICS columns:
naics_code              0
naics_sector_standard    0
dtype: int64

Rows with missing NAICS codes:
```

account_status	ownership_type	date_business_start	date_account_creation	naics_sector	naics_code	naics_description

1.7 Final Validation

The final dataset was validated to ensure:

- No missing values in model features
- Correct datatypes for all predictors
- Variables structure meets modeling requirements
- No duplicate or corrupted records remained
- Sparsity remained below 10%

Inspect final DF

```
# Check for missing values in final DF
missing = df.isnull().sum().sort_values(ascending=False)
missing_percent = (missing / len(df) * 100).round(2)
pd.DataFrame({'Missing': missing, 'Percent': missing_percent})
```

	Missing	Percent
account_status	0	0.0
ownership_type	0	0.0
date_business_start	0	0.0
date_account_creation	0	0.0
naics_sector	0	0.0
naics_code	0	0.0
naics_description	0	0.0
address_city	0	0.0
address_state	0	0.0
lat	0	0.0
lng	0	0.0
zip5	0	0.0
population	0	0.0

2. Create Derived Variables

2.1 Business Age. Calculated from the account creation date

```
if 'date_account_creation' in df.columns:

    df['date_account_creation'] = pd.to_datetime(df['date_account_creation'],
errors='coerce')

    today = pd.Timestamp('today')

    df['business_age'] = (today - df['date_account_creation']).dt.days / 365.25
```

Note: Business age outliers may be retained, as they may reflect actual business behavior. When needed for modeling, transforming the age variable rather than dropping it may be considered.

2.2 Business Age Group. Categorized to support survival modeling.

```
# ======================================

# Create age groups

# ======================================

bins = [0, 1, 3, 5, 10, 20, df['business_age'].max()+1]

labels = ['0-1','1-3','3-5','5-10','10-20','20+']

df['age_group'] = pd.cut(df['business_age'], bins=bins, labels=labels,
right=False)
```

2.3 Business Density. Number of active businesses per ZIP divided by ZIP population times 1000 residents:

```
# ======================================

# Active Business Rate per ZIP

# ======================================

# Filter only active businesses

active_business_df = df[df['business_status'] == 'ACTIVE']

# Count active businesses per ZIP
```

```
zip_counts_active =
active_business_df['zip5'].value_counts().rename_axis('zip5').reset_index(name='bu
siness_count')

# Merge with population

zip_rate_active = zip_pop.reset_index().merge(zip_counts_active, on='zip5',
how='left')

# Fill ZIPs with no active businesses with 0

zip_rate_active['business_count'] = zip_rate_active['business_count'].fillna(0)

# Compute business rate per 1,000 residents

zip_rate_active['business_rate_per_1000'] = zip_rate_active['business_count'] /
zip_rate_active['population'] * 1000
```

2.4 The Active/Closed Rate was created by categorical variables. Supporting descriptive and predictive analysis. Here is an Active/Closed Rate example by ownership type:

```
if 'ownership_type' in df.columns:

    ownership_summary = df.groupby(['ownership_type',
'business_status']).size().unstack(fill_value=0)

    ownership_summary = ownership_summary.rename(columns={0: 'CLOSED', 1:
'ACTIVE'})

    ownership_percentage = ownership_summary.div(ownership_summary.sum(axis=1),
axis=0) * 100
```

2.5 Encoding Categorical Variables. Categorical attributes (ownership type and NAICS sector) were encoded using one-hot encoding to facilitate logistic regression modeling.

```
# ======================================

# Define features

# ======================================

features_num = ['business_age', 'population', 'lat', 'lng', 'business_density']

features_cat = ['ownership_type', 'naics_sector_standard', 'address_city']

target = 'business_status_num'  # 1=Active, 0=Closed
```

```
# ======================================
# Preprocessing pipelines
# ======================================
num_transformer = StandardScaler()
cat_transformer = OneHotEncoder(handle_unknown='ignore', sparse_output=False)
preprocessor = ColumnTransformer(
    transformers=[('num', num_transformer, features_num),
                  ('cat', cat_transformer, features_cat)  ])
```

2.6 Data Preparation for Modeling

- **Checked VIF for multicollinearity (statsmodels)**

```
# ======================================
# Define features
# ======================================
features_num = ['business_age', 'population', 'lat', 'lng', 'business_density']
features_cat = ['ownership_type', 'naics_sector_standard', 'address_city']
target = 'business_status_num'  # 1=Active, 0=Closed
```

```
VIF values for numeric features:
```

	feature	VIF
0	business_age	1.006964
1	population	1.393318
2	lat	1.109734
3	lng	1.195722
4	business_density	1.340950

***No problematic multicollinearity (VIF < 5) in numerical variables**

- **Applied SMOTE for class balancing**
- **Split data into train/test sets**

```
Original training shape: (110251, 8)
Training shape after preprocessing: (110251, 75)
Training shape after SMOTE: (139996, 75)
Test shape after preprocessing: (27563, 75)

Class distribution before SMOTE: Counter({0: 69998, 1: 40253})
Class distribution after SMOTE: Counter({0: 69998, 1: 69998})
```

3. Tools and Techniques Used

Tools:

- **Python** for extraction and preprocessing
- **Pandas** for data manipulation
- **GeoPandas** for spatial processing
- **NumPy** for numerical operations
- **Colab Notebook** for iterative analysis
- **TIGER/Line Shapefiles** for accurate geospatial assignment

Techniques:

- Bounding-box data-quality filtering
- Spatial joins (county and ZCTA layers)
- Feature engineering
- Missing-value handling
- Dummy-variable encoding
- Outlier evaluation
- ZIP-level aggregation
- Geospatial validation and cleanup

3.1 Justification of Tools and Techniques

Advantages:

High precision and reproducibility.
Using Pandas and GeoPandas allows all cleaning, transformations, spatial operations, and preprocessing steps to be implemented programmatically. This ensures repeatability, transparency, and auditability, minimizing human error, which are critical requirements for predictive modeling (Querio.ai, 2025; IBM, n.d.).

Disadvantages:

Environment and version dependency.
Different versions of GeoPandas, Shapely, or Fiona may behave inconsistently, causing errors in spatial joins or coordinate parsing unless the environment is tightly controlled (Querio.ai, 2025).

How Challenges Were Overcome

A fixed Python environment and library versions were used to stabilize spatial operations (Querio.ai, 2025). Bounding-box filters prevented geometry errors during shapefile joins. Mismatched and unknown ZIP codes were identified, logged, and removed. Missing or inconsistent fields were corrected or filtered systematically using the Pandas library (IBM, n.d.). These steps ensured a clean, accurate, fully model-ready dataset.

These steps ensured a clean, accurate, fully model-ready dataset.

Analysis

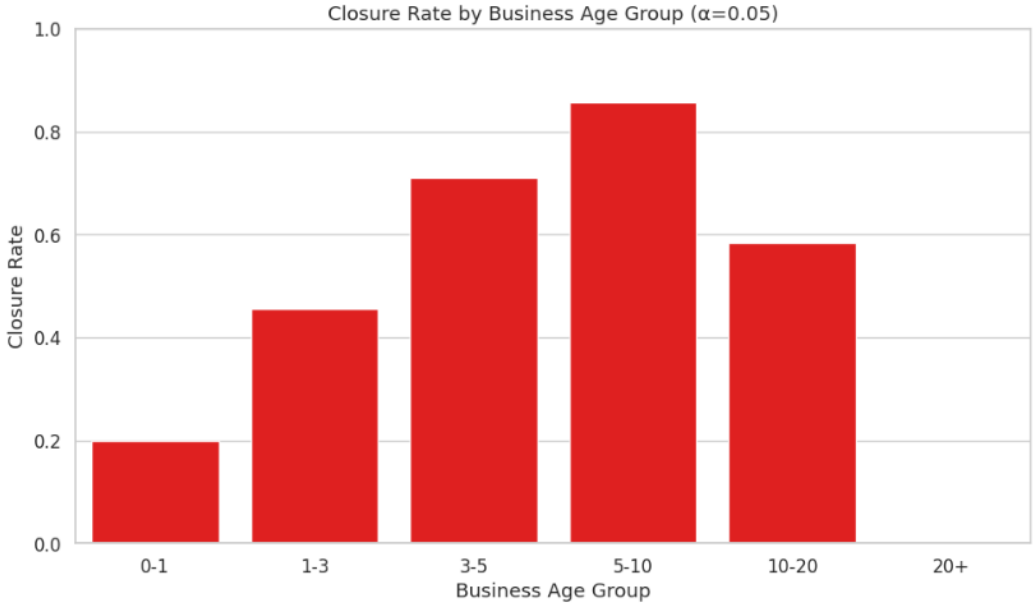
1. Chi-Square Tests for Categorical Variables

Advantage: Easy to apply and interpret when comparing frequencies across categories. The Chi-Square test is a non-parametric test that requires no assumptions about normality, making it well-suited for determining whether two categorical variables are associated.

Disadvantage: Requires sufficiently large, expected cell counts for validity. If sample sizes are small or some categories have very low frequencies, the test becomes unreliable and may result in inflated Type I or Type II errors.

1.1 Closure Rate by Business Age Group

```
Hypothesis Test: Business Age Group vs Closure Rate
H0: Closure rate is independent of business age group
Ha: Closure rate depends on business age group
Significance level: α = 0.05
Chi-Square Statistic = 43982.34
p-value = 0.00000
Conclusion: Reject H0 → Closure rate depends on business age group
```

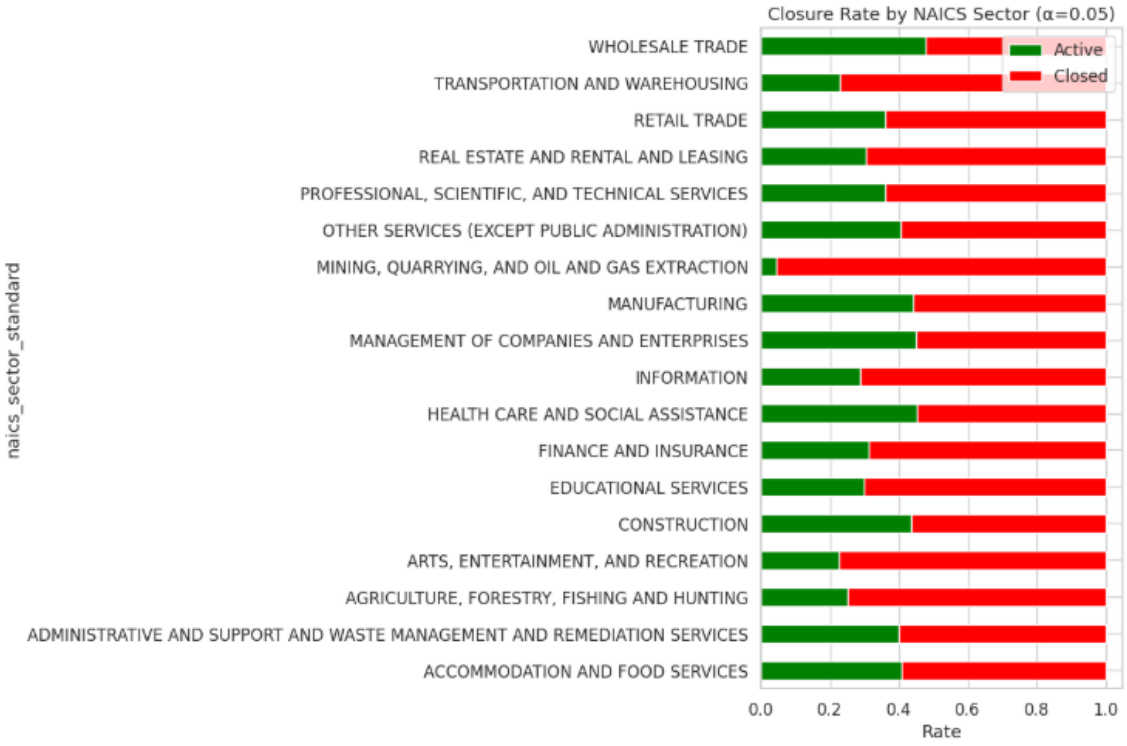


Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ Closure rate *depends* on the business age group.

1.2 Closure Rate by NAICS Sector

```
Hypothesis Test: Business Closure Rate vs NAICS Sector
H0: Closure rate is independent of NAICS sector
Ha: Closure rate depends on NAICS sector
Significance level: α = 0.05
Chi-Square Statistic = 2564.84
p-value = 0.00000
Conclusion: Reject H0 → Closure rate depends on NAICS sector
```

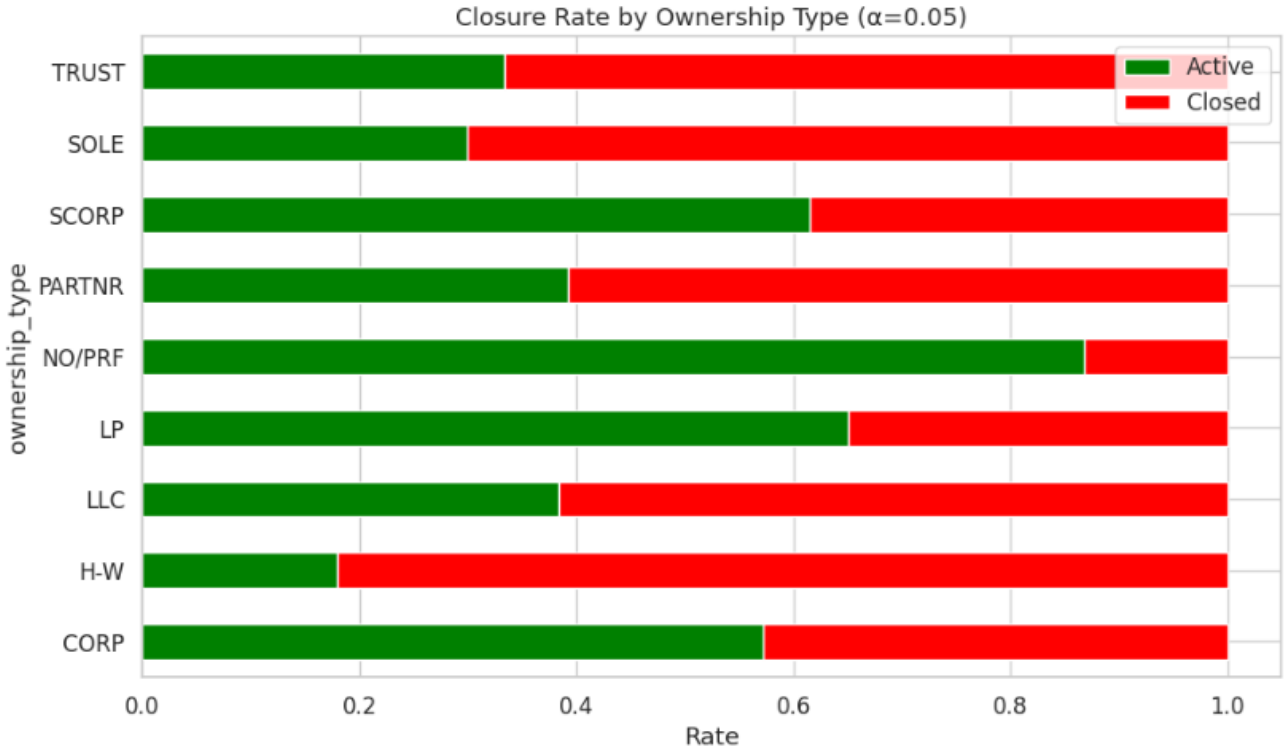


Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ Closure rate *depends* on NAICS Sector.

1.3 Closure Rate by Ownership Type

```
Hypothesis Test: Business Closure Rate vs Ownership Type
H0: Closure rate is independent of ownership type
Ha: Closure rate depends on ownership type
Significance level: α = 0.05
Chi-Square Statistic = 9750.72
p-value = 0.00000
Conclusion: Reject H0 → Closure rate depends on ownership type
```

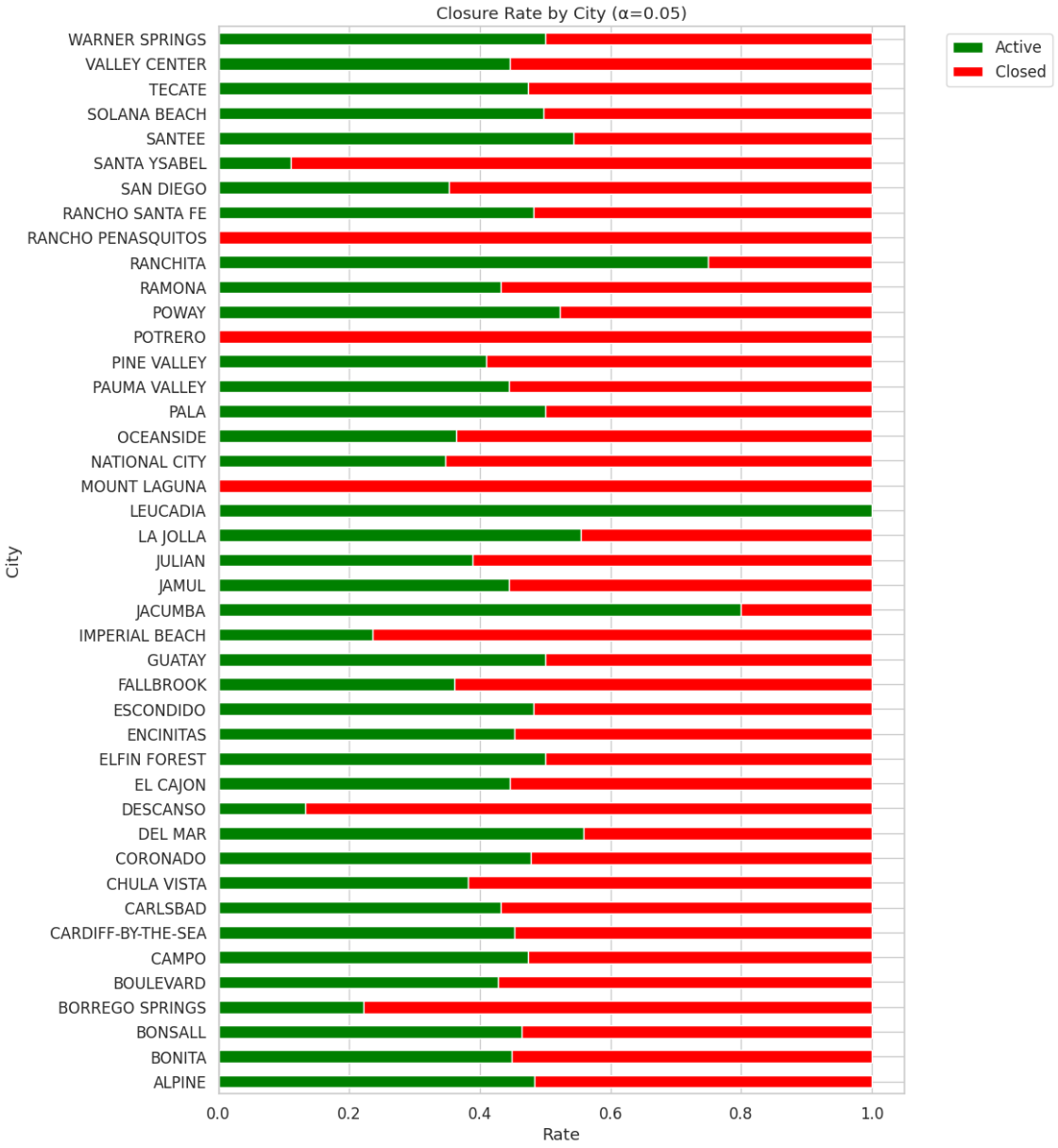


Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ Closure rate *depends* on Ownership Type.

1.4 Closure Rate by City

```
Hypothesis Test: Business Closure Rate vs City
H0: Closure rate is independent of city
Ha: Closure rate depends on city
Significance level: α = 0.05
Chi-Square Statistic = 1106.01
p-value = 0.00000
Conclusion: Reject H0 → Closure rate depends on city
```

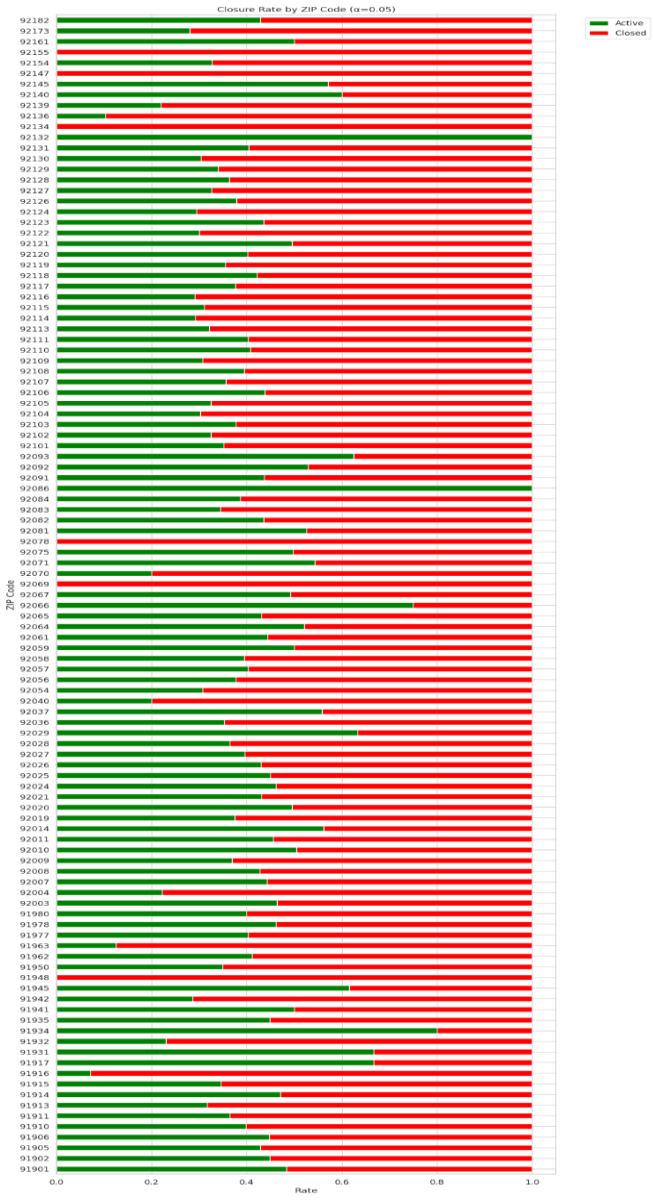


Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ Closure rate *depends* on City.

1.5 Closure Rate by ZIP Code

```
Hypothesis Test: Business Closure Rate vs ZIP Code
H0: Closure rate is independent of ZIP code
Ha: Closure rate depends on ZIP code
Significance level: α = 0.05
Chi-Square Statistic = 2617.48
p-value = 0.00000
Conclusion: Reject H0 → Closure rate depends on ZIP code
```

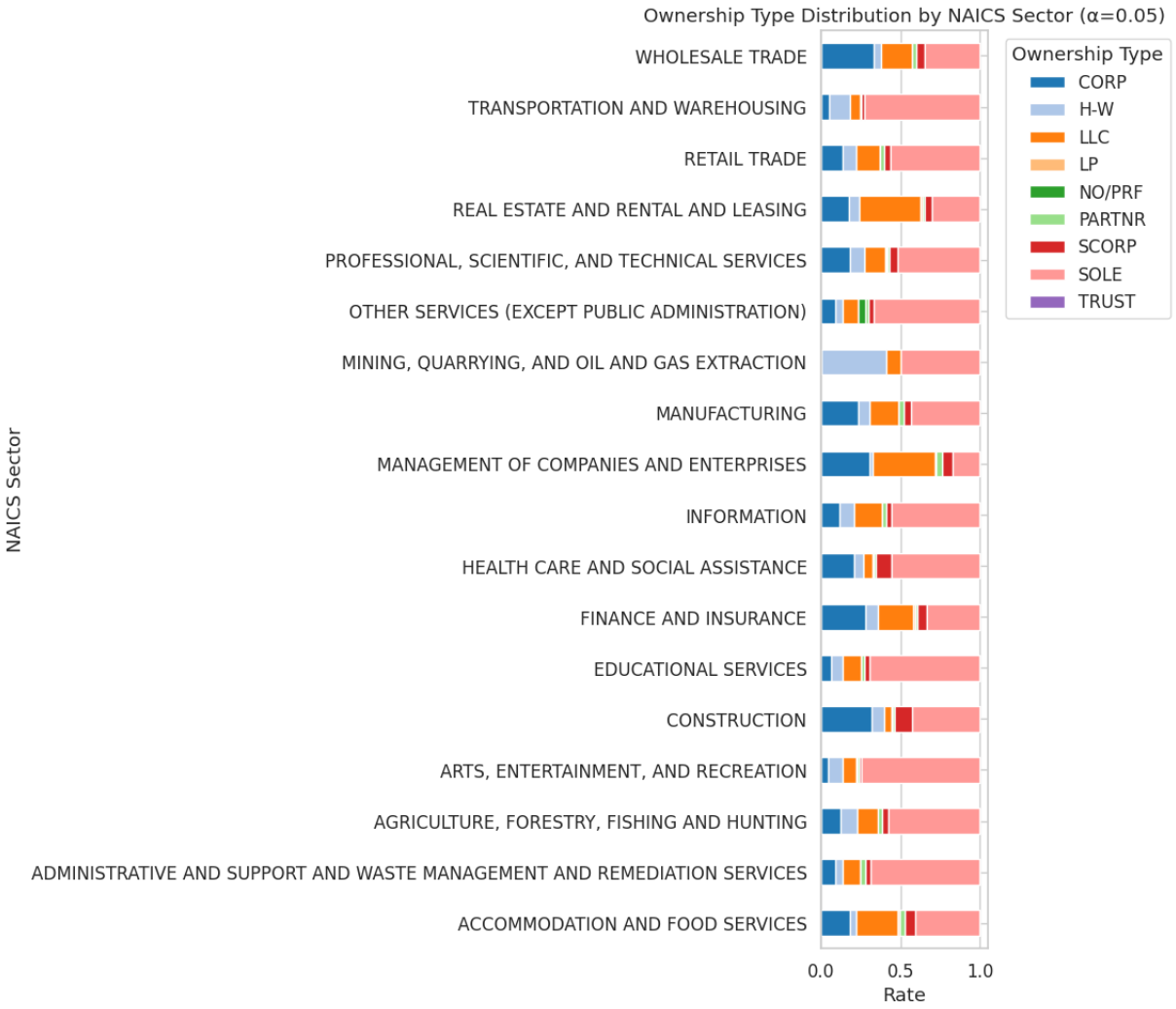


Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ Closure rate *depends* on ZIP Code.

1.6 Ownership Type by NAICS Sector

```
Hypothesis Test: Ownership Type vs NAICS Sector
H0: Ownership type is independent of NAICS sector
Ha: Ownership type depends on NAICS sector
Significance level: α = 0.05
Chi-Square Statistic = 23824.13
p-value = 0.00000
Conclusion: Reject H0 → Ownership type depends on NAICS sector
```



Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ Ownership Type *depends* on NAICS Sector.

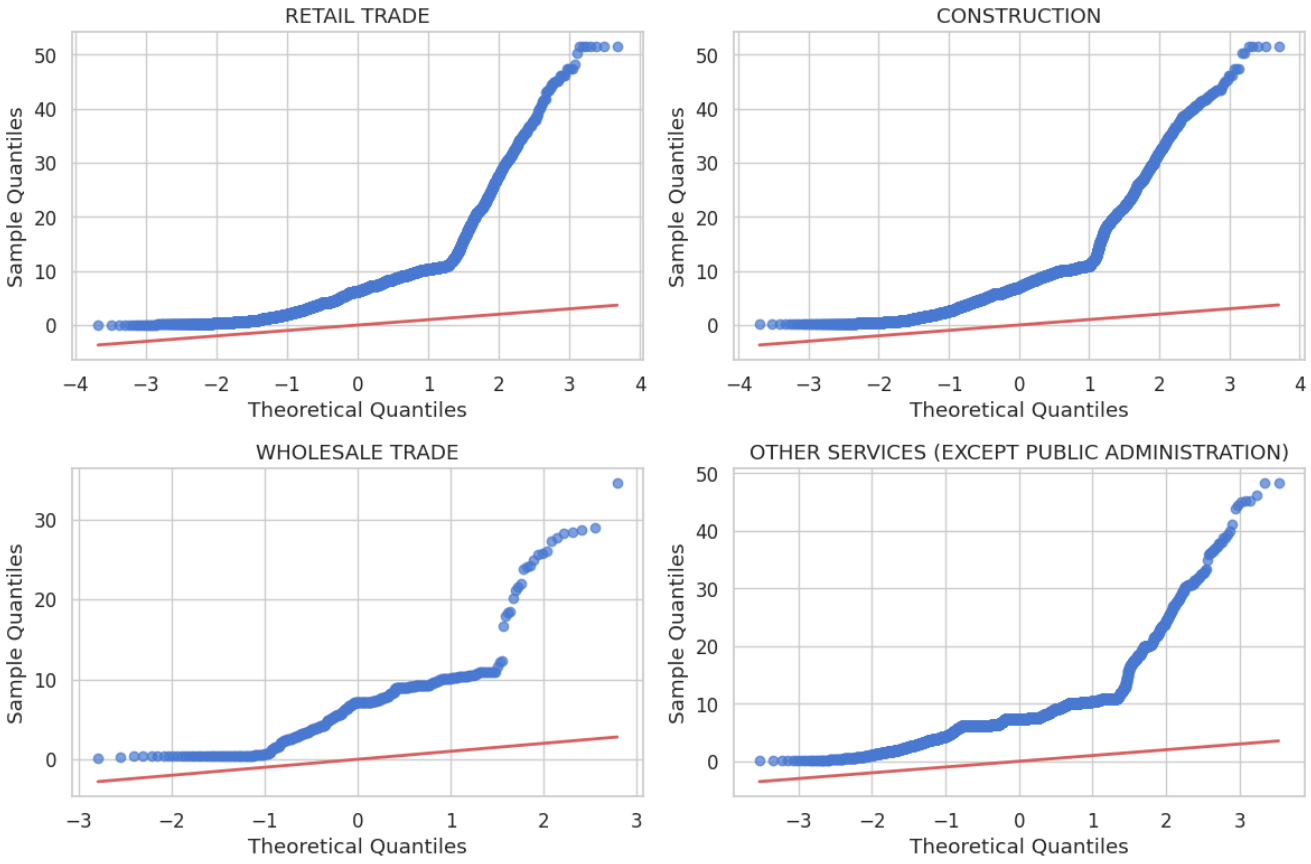
2. Anova/Kruskal-Wallis/Correlation Tests for Quantitative Variables

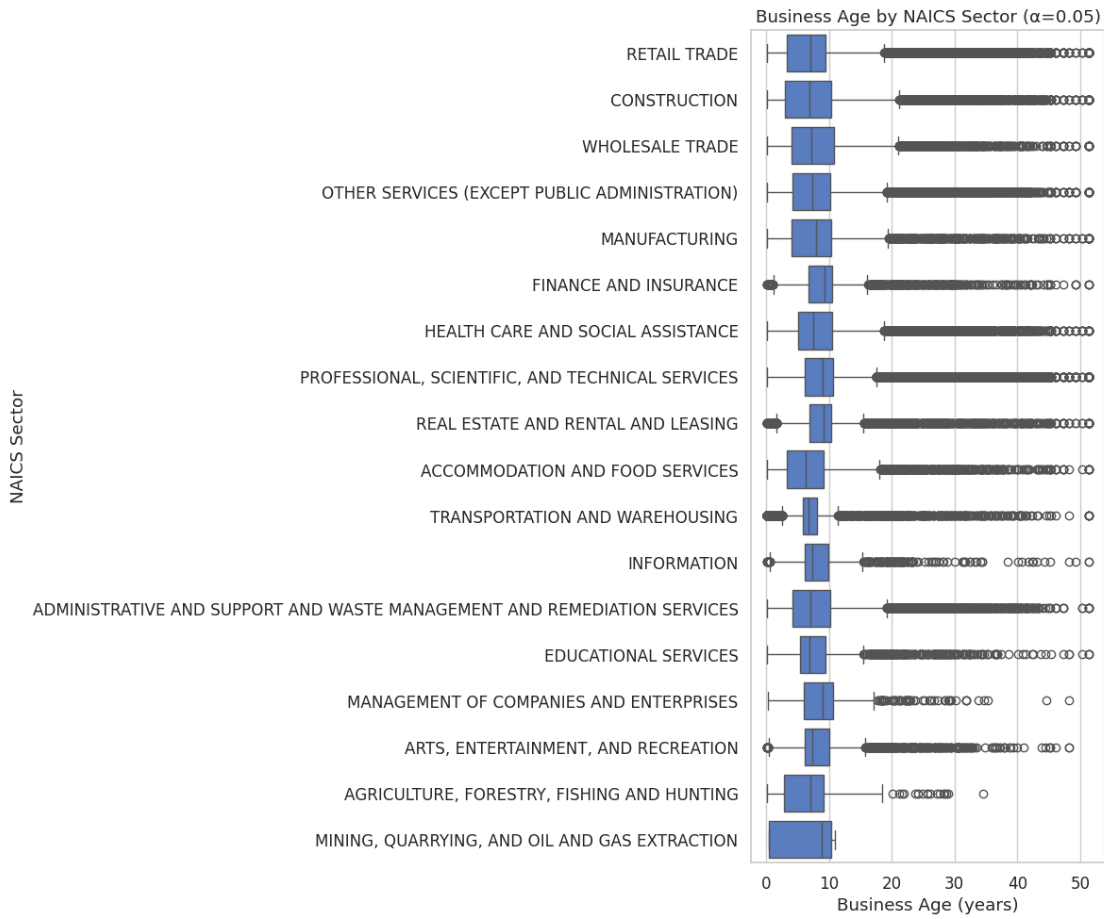
Test	Advantage	Disadvantage
Shapiro–Wilk	Highly sensitive for detecting deviations from normality. The Shapiro–Wilk test is one of the most powerful normality tests, effective even with small sample sizes.	Very sensitive with large samples, often flags normal data as "not normal." With large datasets, even minor, insignificant deviations from normality can produce significant p-values, making interpretation challenging.
Kolmogorov–Smirnov	Works with very large samples (good for datasets > 5,000 rows). Fast and simple to compute. Can test any distribution, not just normality.	Less accurate for normality than the Shapiro–Wilk. Requires estimating the mean and SD first, which can cause bias. Overly sensitive with big datasets (rejects normality even for tiny deviations).
Q-Q Plot	Intuitive visual assessment. It can quickly see deviations from normality, skewness, or heavy tails; patterns are easy to interpret.	Subjective. Interpretation depends on the viewer; small deviations may be hard to quantify, and it doesn't provide a formal p-value like Shapiro-Wilk or Kolmogorov-Smirnov tests.
Levene	Robust test for detecting differences in group variances. Levene's test works well even when data are not normally distributed, making it useful before ANOVA or other parametric tests.	Sensitive to outliers and unequal sample sizes. Extreme values or very small groups can distort the results, potentially suggesting unequal variances when differences are minor.
ANOVA	Powerful for detecting differences in mean values across multiple groups. ANOVA enables the simultaneous comparison of three or more groups, thereby eliminating the need for repeated pairwise tests and controlling Type I error.	Assumes normality and equal variances across groups. If these assumptions are violated, ANOVA results may be invalid unless transformations or alternative methods are used.
Kruskal-Wallis	Does not assume normality, making it robust for skewed or non-normal data. Kruskal–Wallis uses ranks instead of raw values, making it well-suited for distributions that differ significantly from the normal distribution.	Less powerful than ANOVA when assumptions are met. Because it ignores the magnitude of differences (using ranks), it may fail to detect real effects that ANOVA would catch.
Pearson Correlation	Measures the strength and direction of linear relationships with high precision. It uses actual numeric values, making it sensitive and powerful when the relationship is truly linear.	Requires normality and linearity assumptions. If variables are skewed, contain outliers, or the relationship is non-linear, Pearson correlation can be misleading.

Spearman Correlation	Does not assume normality or linearity, making it robust to skewed data and outliers. Spearman detects monotonic relationships, even when the pattern is curved or uneven.	Less sensitive than Pearson when the relationship is truly linear. Because it uses ranks instead of raw values, it can underestimate the strength of linear associations.

2.1 Business Age by NAICS Sector

```
Hypothesis Test: Business Age vs NAICS Sector
H0: Mean/median business age is the same across sectors
Ha: At least one sector has a different mean/median business age
Significance level: α = 0.05
Normality test (Shapiro-Wilk/K-S) passed for all groups: False
Levene test for equal variances: stat=169.99, p=0.00000 → Unequal variance
H-statistic = 7318.76
p-value = 0.00000
Conclusion: Reject H0 → At least one sector has a different business age
Note: Normality or equal variance violated → Kruskal-Wallis used
```





Output observed:

- **$p < 0.05$ → Reject H_0**
  ⇒ Business age *depends* on the NAICS sector.

<u>Post-hoc test results:</u>

Post-hoc test results:

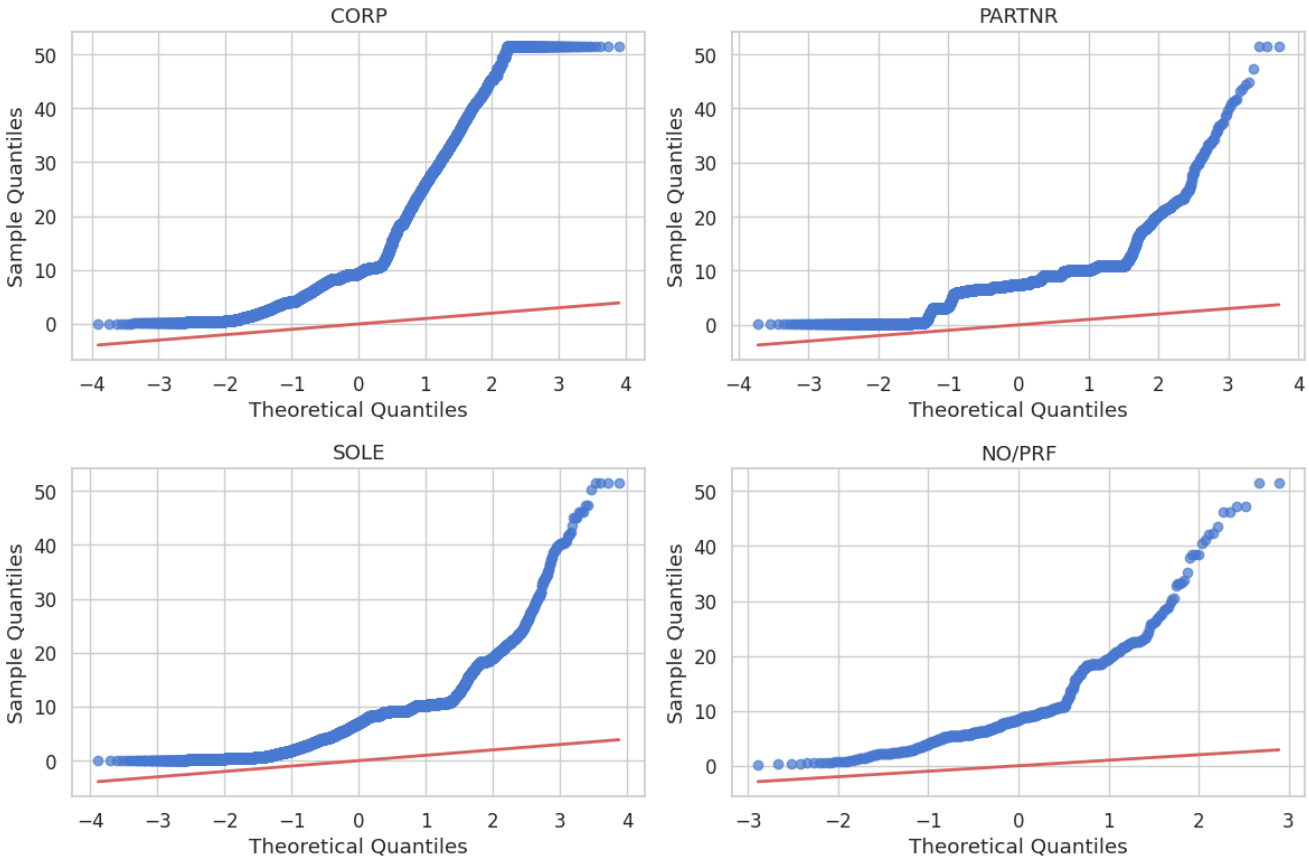
	ACCOMMODATION AND FOOD SERVICES	ADMINISTRATIVE AND SUPPORT AND WASTE MANAGEMENT AND REMEDIATION SERVICES	AGRICULTURE, FORESTRY, FISHING AND HUNTING	ARTS, ENTERTAINMENT, AND RECREATION	CONSTRUCTION	EDUCATIONAL SERVICES
ACCOMMODATION AND FOOD SERVICES	1.000000e+00	3.267858e-44	1.000000e+00	6.406788e-48	9.892716e-20	4.759399e-17
ADMINISTRATIVE AND SUPPORT AND WASTE MANAGEMENT AND REMEDIATION SERVICES	3.267858e-44	1.000000e+00	8.719487e-03	4.018952e-01	2.000508e-02	1.000000e+00
AGRICULTURE, FORESTRY, FISHING AND HUNTING	1.000000e+00	8.719487e-03	1.000000e+00	1.123935e-04	6.454848e-01	3.597022e-01
ARTS, ENTERTAINMENT, AND RECREATION	6.406788e-48	4.018952e-01	1.123935e-04	1.000000e+00	1.882792e-07	1.709819e-04
CONSTRUCTION	9.892716e-20	2.000508e-02	6.454848e-01	1.882792e-07	1.000000e+00	1.000000e+00
EDUCATIONAL SERVICES	4.759399e-17	1.000000e+00	3.597022e-01	1.709819e-04	1.000000e+00	1.000000e+00
FINANCE AND INSURANCE	3.078583e-189	1.447389e-90	2.229227e-32	1.089463e-59	1.208139e-108	1.298383e-89
HEALTH CARE AND SOCIAL ASSISTANCE	6.199424e-118	2.434484e-18	5.237874e-09	4.058452e-04	1.745157e-33	1.127423e-21
INFORMATION	4.224451e-30	1.480011e-02	4.815275e-08	1.000000e+00	2.808315e-07	2.038948e-05
MANAGEMENT OF COMPANIES AND ENTERPRISES	4.401038e-29	2.842838e-11	4.801438e-14	9.518481e-08	2.573470e-15	6.211400e-14
MANUFACTURING	2.589770e-39	3.278121e-02	1.889524e-05	1.000000e+00	5.451388e-08	1.881597e-05
MINING, QUARRYING, AND OIL AND GAS EXTRACTION	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
OTHER SERVICES (EXCEPT PUBLIC ADMINISTRATION)	3.245441e-101	2.819702e-05	1.520307e-05	1.000000e+00	1.301583e-17	1.423287e-09
PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	0.000000e+00	3.544885e-175	1.480004e-24	1.539725e-74	1.041952e-197	4.918738e-125
REAL ESTATE AND RENTAL AND LEASING	2.484748e-291	1.174801e-118	8.477227e-25	4.407905e-80	3.952384e-142	9.408380e-101
RETAIL TRADE	8.484185e-21	8.429027e-10	1.000000e+00	3.909087e-16	1.000000e+00	1.000000e+00
TRANSPORTATION AND WAREHOUSING	7.007914e-02	2.577712e-29	1.000000e+00	2.288130e-34	1.110810e-09	1.150089e-08
WHOLESALE TRADE	1.481704e-46	1.721128e-03	4.888737e-08	1.000000e+00	2.842838e-10	4.277987e-07

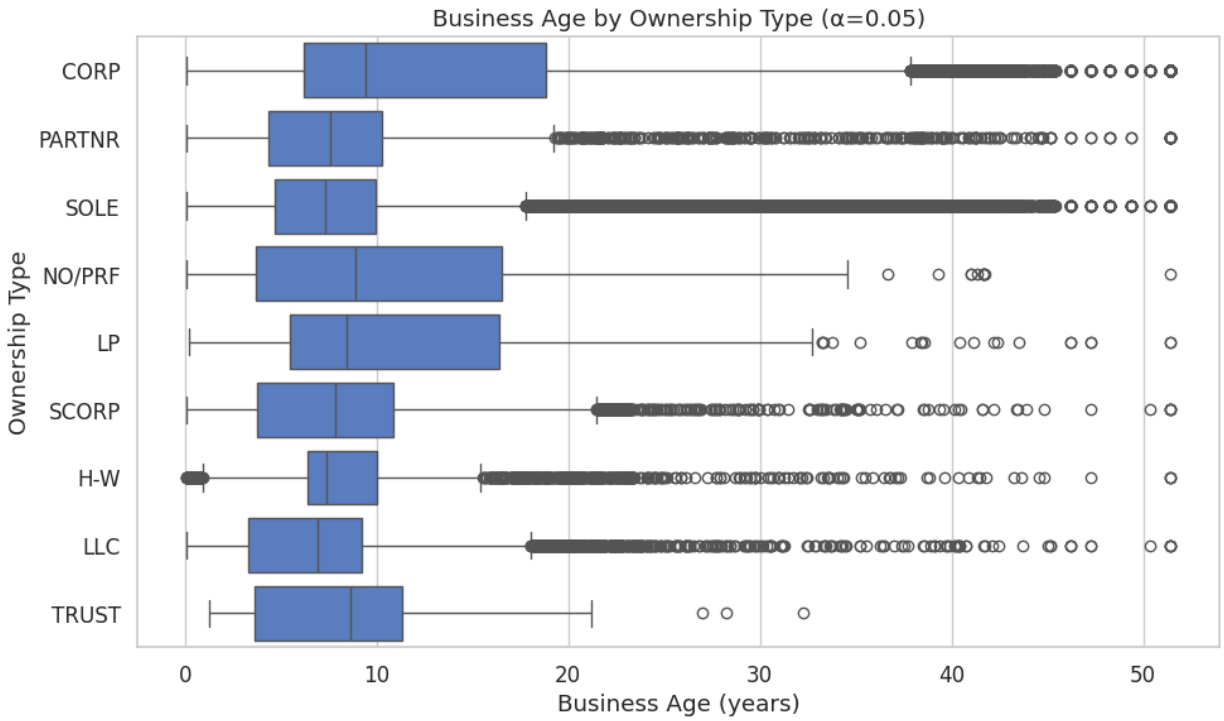
	FINANCE AND INSURANCE	HEALTH CARE AND SOCIAL ASSISTANCE	INFORMATION	MANAGEMENT OF COMPANIES AND ENTERPRISES	MANUFACTURING	MINING, QUARRYING, AND OIL AND GAS EXTRACTION
ACCOMMODATION AND FOOD SERVICES	3.078583e-189	6.199424e-118	4.224451e-30	4.401038e-29	2.589770e-39	1.0
ADMINISTRATIVE AND SUPPORT AND WASTE MANAGEMENT AND REMEDIATION SERVICES	1.447389e-90	2.434484e-18	1.480011e-02	2.842838e-11	3.278121e-02	1.0
AGRICULTURE, FORESTRY, FISHING AND HUNTING	2.229227e-32	5.237874e-09	4.815275e-08	4.801438e-14	1.889524e-05	1.0
ARTS, ENTERTAINMENT, AND RECREATION	1.089463e-59	4.058452e-04	1.000000e+00	9.518481e-08	1.000000e+00	1.0
CONSTRUCTION	1.208139e-108	1.745157e-33	2.808315e-07	2.573470e-15	5.451388e-08	1.0
EDUCATIONAL SERVICES	1.298383e-89	1.127423e-21	2.038948e-05	6.211400e-14	1.881597e-05	1.0
FINANCE AND INSURANCE	1.000000e+00	5.256982e-45	1.588090e-27	1.000000e+00	7.800900e-42	1.0
HEALTH CARE AND SOCIAL ASSISTANCE	5.256982e-45	1.000000e+00	1.000000e+00	5.351112e-04	1.000000e+00	1.0
INFORMATION	1.588090e-27	1.000000e+00	1.000000e+00	2.192992e-04	1.000000e+00	1.0
MANAGEMENT OF COMPANIES AND ENTERPRISES	1.000000e+00	5.351112e-04	2.192992e-04	1.000000e+00	5.872504e-08	1.0
MANUFACTURING	7.800900e-42	1.000000e+00	1.000000e+00	5.872504e-08	1.000000e+00	1.0
MINING, QUARRYING, AND OIL AND GAS EXTRACTION	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.0
OTHER SERVICES (EXCEPT PUBLIC ADMINISTRATION)	3.911170e-75	2.802283e-08	1.000000e+00	1.325525e-07	1.000000e+00	1.0
PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	1.009203e-08	1.408582e-82	1.050492e-18	1.000000e+00	1.058348e-37	1.0
REAL ESTATE AND RENTAL AND LEASING	2.800110e-04	1.581453e-44	2.559021e-18	1.000000e+00	4.959957e-34	1.0
RETAIL TRADE	7.118095e-143	1.169830e-87	1.819877e-11	5.414488e-18	2.854780e-14	1.0
TRANSPORTATION AND WAREHOUSING	4.165479e-171	1.980519e-98	5.089895e-22	1.898884e-24	1.830115e-28	1.0
WHOLESALE TRADE	3.945514e-42	1.000000e+00	1.000000e+00	1.314048e-05	1.000000e+00	1.0

	OTHER SERVICES (EXCEPT PUBLIC ADMINISTRATION)	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	REAL ESTATE AND RENTAL AND LEASING	RETAIL TRADE	TRANSPORTATION AND WAREHOUSING	WHOLESALE TRADE
ACCOMMODATION AND FOOD SERVICES	3.245441e-101	0.000000e+00	2.484748e-291	8.484185e-21	7.007914e-02	1.481704e-46
ADMINISTRATIVE AND SUPPORT AND WASTE MANAGEMENT AND REMEDIATION SERVICES	2.819702e-05	3.544885e-175	1.174801e-118	8.429027e-10	2.577712e-29	1.721128e-03
AGRICULTURE, FORESTRY, FISHING AND HUNTING	1.520307e-05	1.480004e-24	8.477227e-25	1.000000e+00	1.000000e+00	4.888737e-08
ARTS, ENTERTAINMENT, AND RECREATION	1.000000e+00	1.539725e-74	4.407905e-80	3.909087e-16	2.288130e-34	1.000000e+00
CONSTRUCTION	1.301583e-17	1.041952e-197	3.952384e-142	1.000000e+00	1.110810e-09	2.842838e-10
EDUCATIONAL SERVICES	1.423287e-09	4.918738e-125	9.408380e-101	1.000000e+00	1.150089e-08	4.277987e-07
FINANCE AND INSURANCE	3.911170e-75	1.009203e-08	2.800110e-04	7.118095e-143	4.165479e-171	3.945514e-42
HEALTH CARE AND SOCIAL ASSISTANCE	2.802283e-08	1.408582e-82	1.581453e-44	1.169830e-87	1.980519e-98	1.000000e+00
INFORMATION	1.000000e+00	1.050492e-18	2.559021e-18	1.819877e-11	5.089895e-22	1.000000e+00
MANAGEMENT OF COMPANIES AND ENTERPRISES	1.325525e-07	1.000000e+00	1.000000e+00	5.414488e-18	1.898884e-24	1.314048e-05
MANUFACTURING	1.000000e+00	1.058348e-37	4.959957e-34	2.854780e-14	1.830115e-28	1.000000e+00
MINING, QUARRYING, AND OIL AND GAS EXTRACTION	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
OTHER SERVICES (EXCEPT PUBLIC ADMINISTRATION)	1.000000e+00	2.889094e-188	7.884057e-105	1.138447e-49	2.278018e-81	1.000000e+00
PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	2.889094e-188	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	2.539283e-39
REAL ESTATE AND RENTAL AND LEASING	7.884057e-105	1.000000e+00	1.000000e+00	7.255241e-243	2.831905e-273	7.933784e-35
RETAIL TRADE	1.138447e-49	0.000000e+00	7.255241e-243	1.000000e+00	9.478254e-09	3.144082e-18
TRANSPORTATION AND WAREHOUSING	2.278018e-81	0.000000e+00	2.831905e-273	9.478254e-09	1.000000e+00	1.679950e-34
WHOLESALE TRADE	1.000000e+00	2.539283e-39	7.933784e-35	3.144082e-18	1.679950e-34	1.000000e+00

2.2 Business Age by Ownership Type

```
Hypothesis Test: Business Age vs Ownership Type
H0: Mean/median business age is the same across ownership types
Ha: At least one ownership type has a different mean/median business age
Significance level: α = 0.05
Normality test (Shapiro-Wilk/K-S) passed for all groups: False
Levene test for equal variances: stat=1167.70, p=0.00000 → Unequal variance
H-statistic = 6349.01
p-value = 0.00000
Conclusion: Reject H0 → At least one ownership type has a different mean/median business age
Note: Normality or equal variance violated → Kruskal-Wallis used
```





Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ Business age *depends* on the Ownership Type.

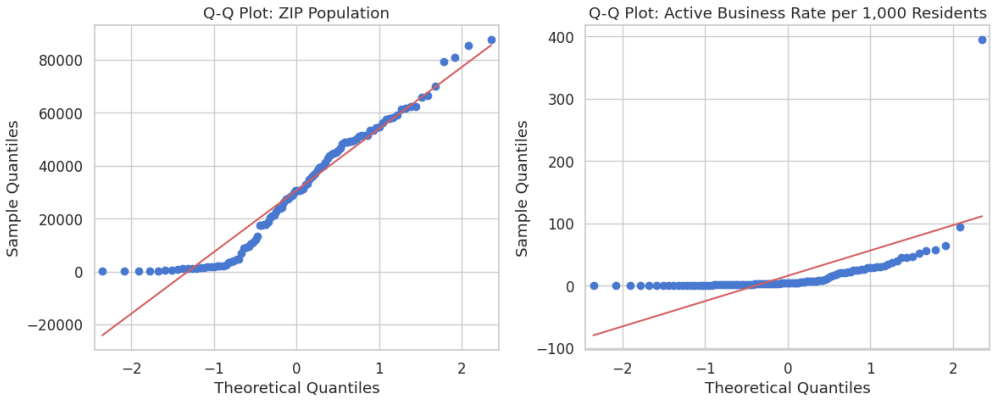
Post-hoc test results:

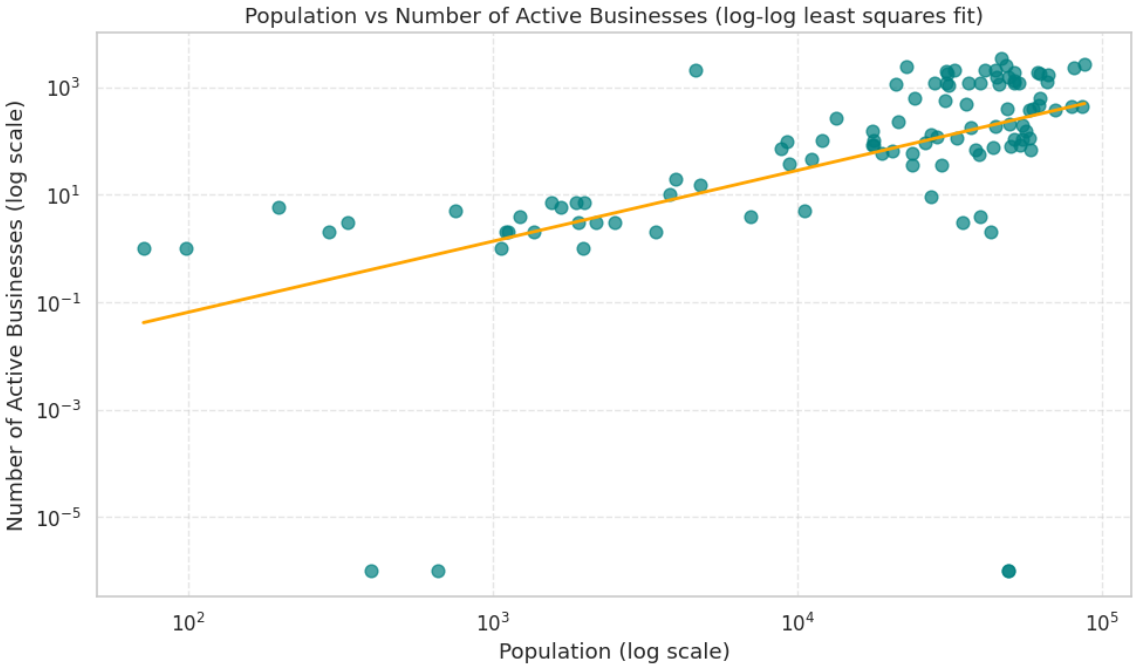
Post-hoc test results:

	CORP	H-W	LLC	LP	NO/PRF	PARTNR	SCORP	SOLE	TRUST
CORP	1.000000e+00	0.000000e+00	0.000000e+00	4.247954e-06	1.399268e-12	2.232462e-104	1.569601e-198	0.000000e+00	1.0
H-W	0.000000e+00	1.000000e+00	3.921585e-52	4.251921e-06	9.511324e-08	1.000000e+00	3.227372e-01	7.890330e-05	1.0
LLC	0.000000e+00	3.921585e-52	1.000000e+00	2.617890e-20	4.673781e-29	1.867619e-28	9.269845e-56	5.069150e-65	1.0
LP	4.247954e-06	4.251921e-06	2.617890e-20	1.000000e+00	1.000000e+00	1.021211e-03	5.991011e-04	2.137729e-09	1.0
NO/PRF	1.399268e-12	9.511324e-08	4.673781e-29	1.000000e+00	1.000000e+00	3.242194e-04	1.108128e-04	8.843778e-13	1.0
PARTNR	2.232462e-104	1.000000e+00	1.867619e-28	1.021211e-03	3.242194e-04	1.000000e+00	1.000000e+00	1.141611e-04	1.0
SCORP	1.569601e-198	3.227372e-01	9.269845e-56	5.991011e-04	1.108128e-04	1.000000e+00	1.000000e+00	7.605767e-11	1.0
SOLE	0.000000e+00	7.890330e-05	5.069150e-65	2.137729e-09	8.843778e-13	1.141611e-04	7.605767e-11	1.000000e+00	1.0
TRUST	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.0

2.3 Active Business Rate per 1,000 Residents by ZIP Code Population

```
Hypothesis Test: Number of Active Business vs Population
Shapiro-Wilk: p = 0.00008 → Not normal
Shapiro-Wilk: p = 0.00000 → Not normal
Method: Spearman
Correlation coefficient = 0.6525735632588089
p-value = 3.480316750606514e-14
Conclusion: Reject H0 → Population and Number of Active Business are significantly correlated
```





Output observed:

- **$p < 0.05 \rightarrow$ Reject H_0**
  ⇒ The number of Active businesses is correlated with the Population.

NOTE: There is a positive association between the population of a ZIP Code and the number of Active Businesses; larger populations are associated with higher business activity.

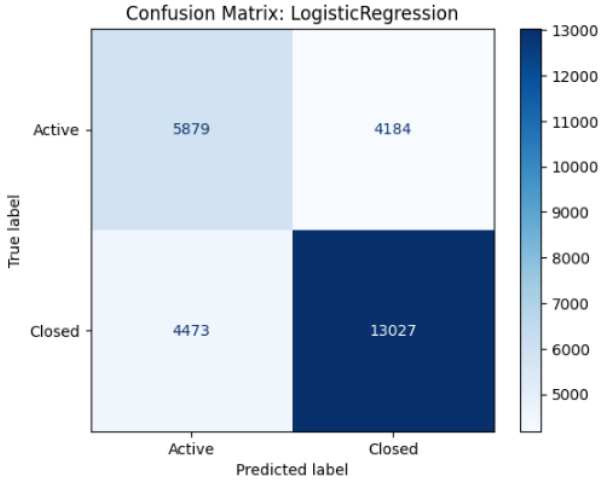
3. Supervised and Unsupervised Machine Learning Models

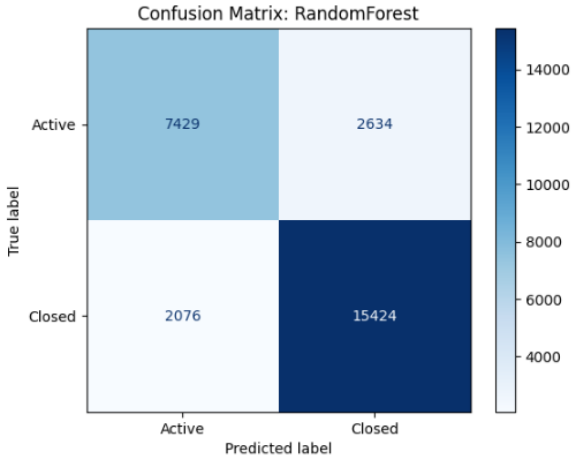
3.1 Supervised: Predictive Models

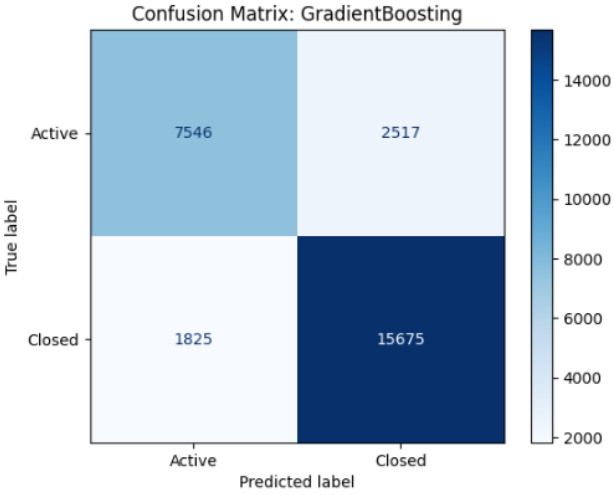
Model	Advantage	Disadvantage
Logistic Regression	High interpretability. Coefficients directly represent the direction and magnitude of relationships between predictors and the probability of closure. Performs well when the relationship between predictors and the log-odds of the outcome is approximately linear.	Poor performance on nonlinear relationships unless significant feature engineering is applied (interaction terms, polynomial features). Assumes independence, linearity in log-odds, and low multicollinearity, all of which can be difficult to satisfy with real-world business datasets.
Random Forest	Handles nonlinear relationships and complex interactions automatically; robust to outliers and noisy data. Offers good baseline accuracy with minimal tuning and provides useful estimates of feature importance.	Less interpretable than logistic regression; the ensemble of many trees makes it difficult to understand exact decision pathways. It can become computationally expensive with very large datasets or many trees.
Gradient Boosting	High predictive accuracy. Boosting sequentially corrects errors from prior trees, leading to strong performance on tabular datasets. It can capture subtle patterns through iterative refinement.	Sensitive to hyperparameters and can overfit if the learning rate or tree depth is not carefully tuned. Training time increases because trees are built sequentially rather than in parallel.
XGBoost	State-of-the-art performance on structured/tabular data; highly optimized, regularized, and fast. Built-in handling for missing values, efficient tree splitting, and strong ability to model nonlinear interactions. Extensive tuning options enable fine-grained control over the bias-variance tradeoff.	Complexity and tuning cost performance depend heavily on correctly setting many hyperparameters (learning rate, depth, subsampling, and regularization). Less interpretable without post-hoc tools like SHAP.
LightGBM	Very fast and scalable, it utilizes histogram-based splitting and leaf-wise growth, making it faster than XGBoost for large datasets. Excellent performance on mixed feature types and high-cardinality categorical variables.	Prone to overfitting due to leaf-wise tree growth if not properly regularized. Sensitive to feature distributions perform differently if the data is not well-prepared or balanced.

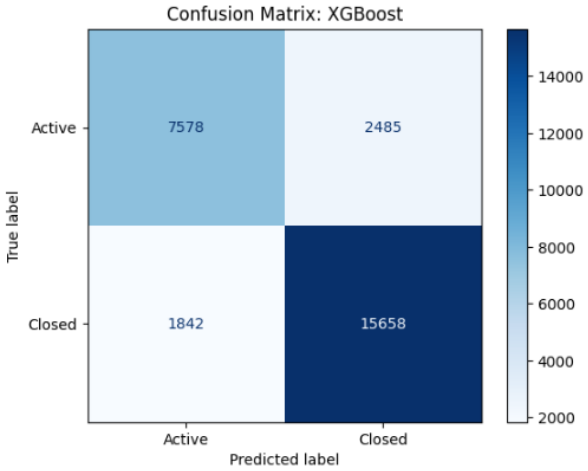
SMOTE	Reduces class imbalance by generating synthetic minority-class samples, which improves model performance, especially recall without simply duplicating data, thereby reducing overfitting compared to traditional oversampling (Nemade et al., 2023)	It can create unrealistic or noisy synthetic samples, especially when minority points are sparse or near class boundaries, which may lead to overfitting, create class overlap, and reduce model reliability if not used carefully. Must be applied only to the training set (Nemade et al., 2023)

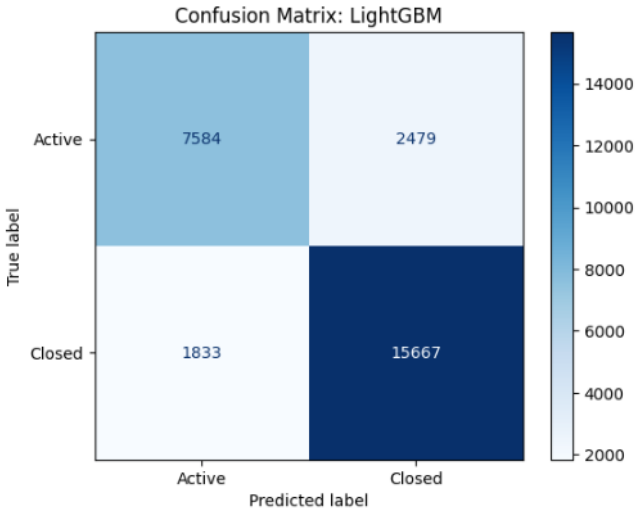
Models' Confusion Matrices:

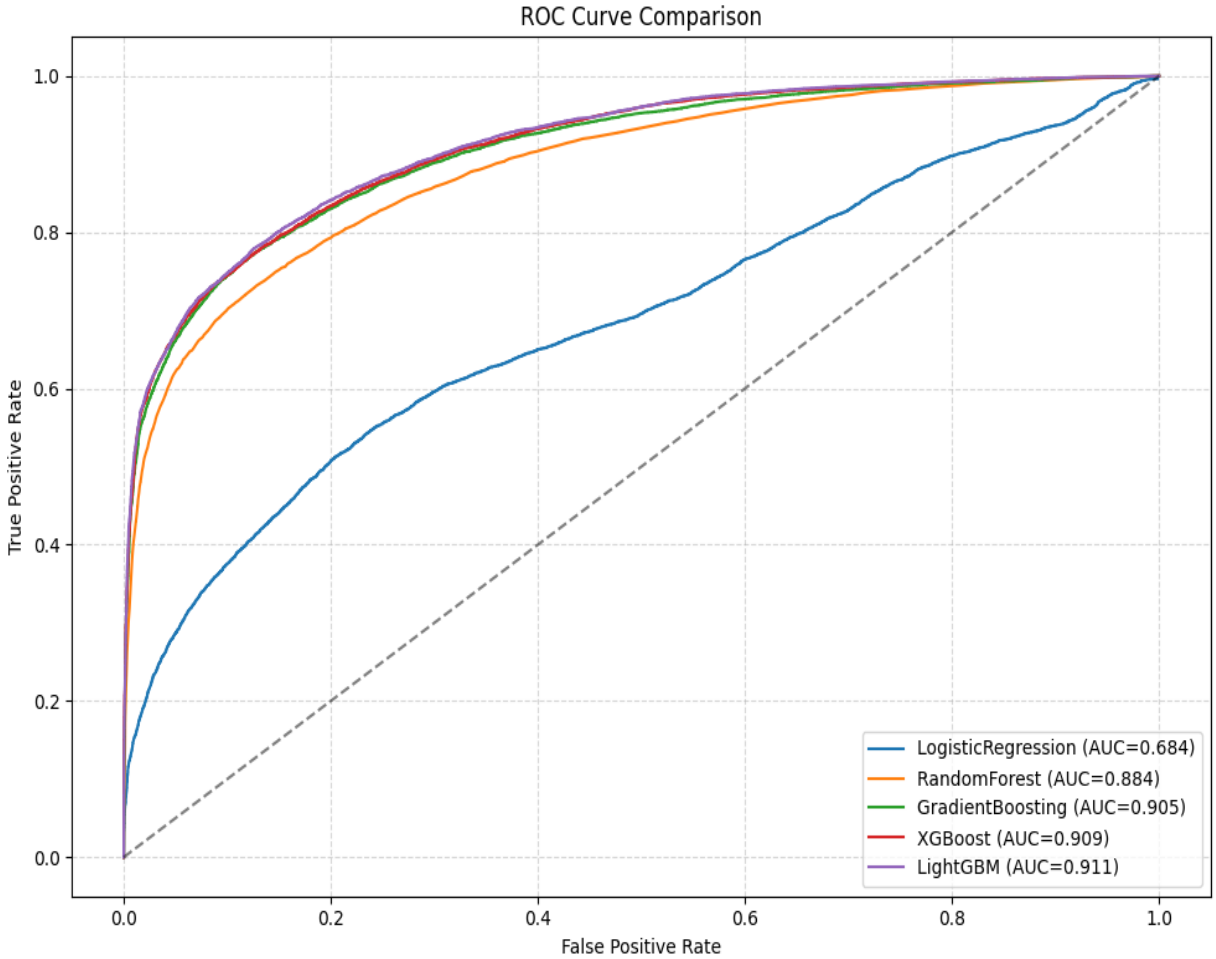









Confusion Matrix: LightGBM
Active
Closed
True label
7584
2479
1833
15667
Active
Closed
Predicted label
14000
12000
10000
8000
6000
4000
2000


ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
LogisticRegression (AUC=0.684)
RandomForest (AUC=0.884)
GradientBoosting (AUC=0.905)
XGBoost (AUC=0.909)
LightGBM (AUC=0.911)

Model Comparison Summary:

	Model	Accuracy	Precision	Recall	F1	ROC_AUC
0	LightGBM	0.835679	0.827757	0.760220	0.792553	0.911295
1	XGBoost	0.834613	0.829190	0.754970	0.790341	0.908955
2	GradientBoosting	0.834506	0.827840	0.756519	0.790574	0.904714
3	RandomForest	0.811549	0.789672	0.740942	0.764531	0.884499
4	LogisticRegression	0.661834	0.592019	0.582236	0.587087	0.683970

Best model based on both Accuracy and ROC-AUC: LightGBM
Metrics:
accuracy: 0.836
precision: 0.828
recall: 0.760
f1: 0.793
roc_auc: 0.911

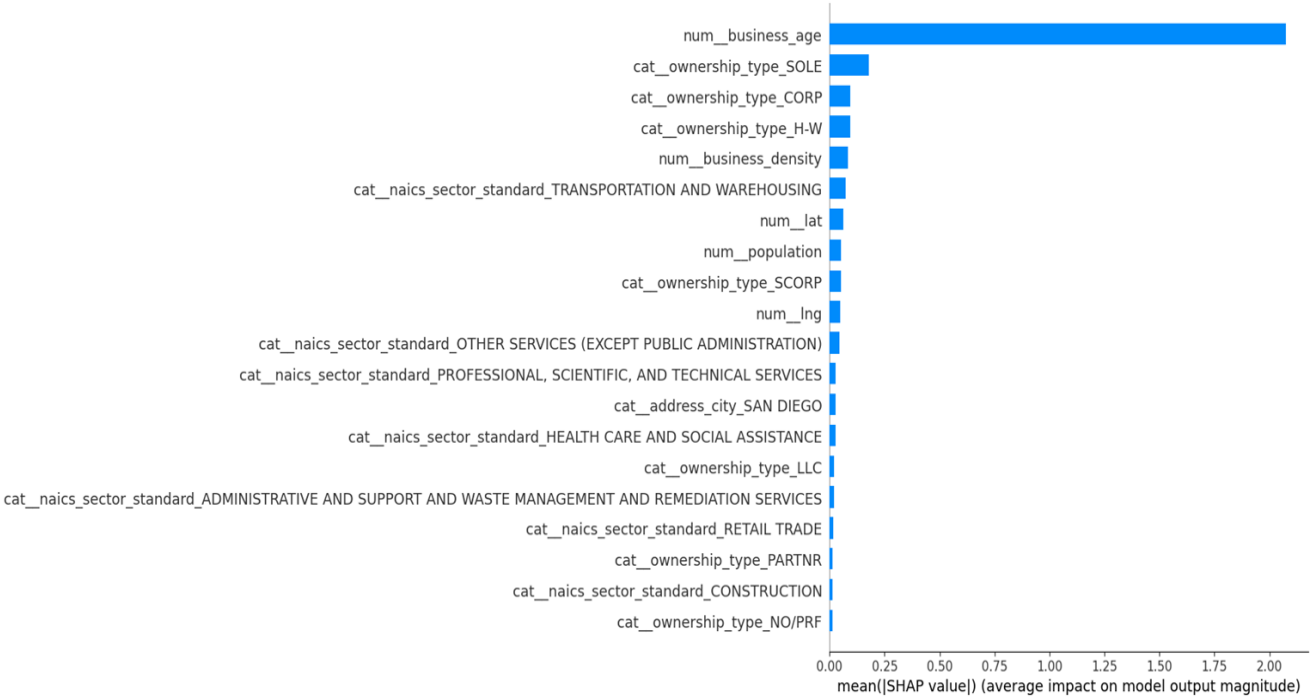
Model Comparison Summary

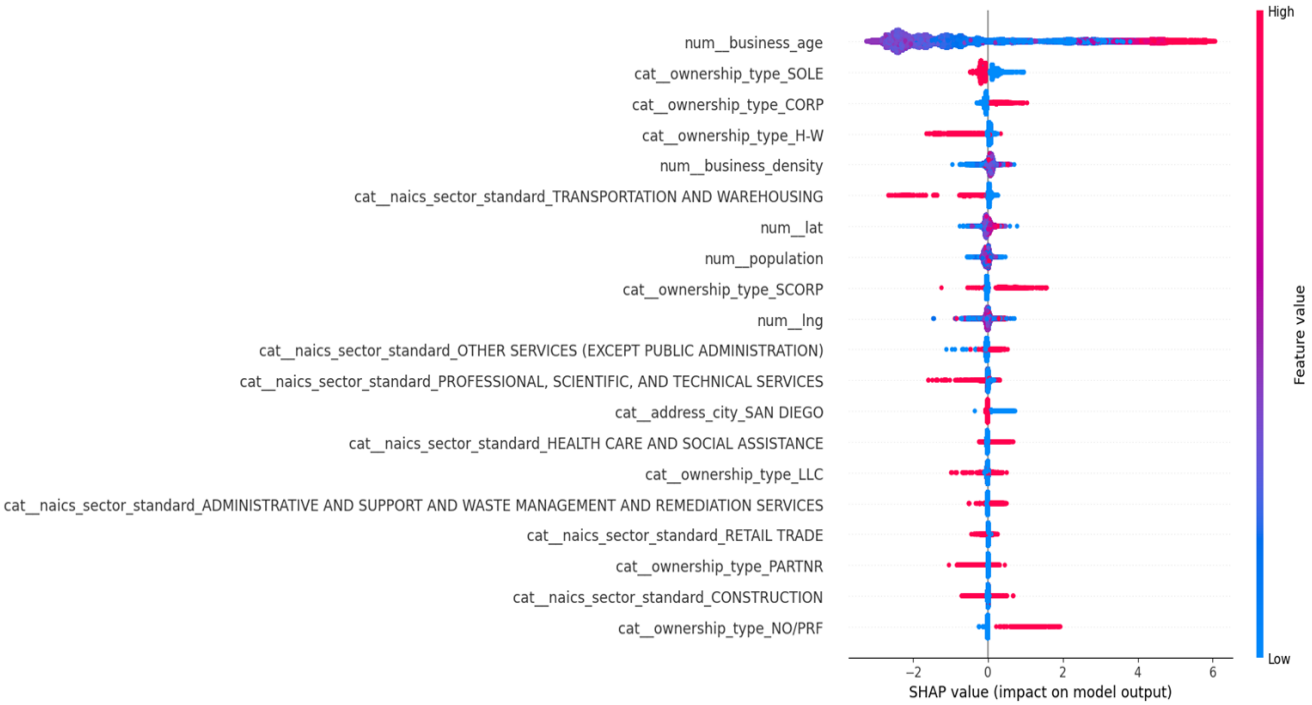
The comparison shows that **LightGBM and XGBoost outperform all other models** across every major metric. LightGBM slightly outperforms the others, achieving the highest accuracy (0.836) and ROC-AUC (0.911), which indicates that it distinguishes well between active and closed businesses while maintaining a strong overall performance. XGBoost yields nearly identical results, and Gradient Boosting also performs competitively, albeit with slightly lower recall and ROC-AUC scores.

The **Random Forest** model performs reasonably well but lags behind the boosting models, reflecting its weaker ability to capture complex patterns in the data.
Logistic Regression performs the worst across all metrics, suggesting that the relationships in the dataset are nonlinear and better captured by tree-based and boosting algorithms.

Overall, **LightGBM is the best-performing model**, achieving an accuracy of 83.6%, an F1 score of 0.793, and an ROC-AUC of 0.911, indicating strong predictive power for business survival. While precision (82.8%) is slightly higher than recall (76.0%), the model effectively balances the rates of false positives and false negatives, making it suitable for practical decision-making.

SHAP Analysis for Model Interpretability





SHAP Feature Importance Report

The SHAP summary plot illustrates the relative importance and directional influence of each predictor on the model's probability of business survival. Each dot represents one business, where:

- **Color** reflects the feature value (red = high, blue = low).
- **Horizontal position** reflects the SHAP value (positive = pushes prediction toward survival; negative = pushes prediction toward non-survival).
- **Vertical order** reflects the overall importance of elements in the model.

This visualization identifies which variables have a meaningful influence on the model's output and which contribute minimally.

Key Findings

1. Business age is the strongest predictor of survival

The variable **num__business_age** is by far the most dominant feature.

- **High business age (red)** → positive SHAP values → **older businesses are more likely to survive**.
- **Low business age (blue)** → negative SHAP values → **younger businesses face higher closure risk.**

Interpretation*:* Established businesses tend to be more resilient (the dominant driver of survival), while very young businesses have higher early-stage closure rates (most vulnerable).

2. Ownership structure meaningfully affects survival, but effects differ by category

Several ownership-type variables appear among the top predictors:

- **cat__ownership_type_SOLE**, **cat__ownership_type_CORP**, **cat__ownership_type_LLC**, etc.

General patterns:

- Some ownership forms, such as SOLE proprietors, exhibit red (high value) production, which tends to push predictions toward non-survival and higher risk.
- Others, such as **CORP and SCORP structures,** often push predictions toward survival, indicating greater stability.

- **LLCs** show a mild positive effect but are less influential than corporate structures.

- **Non-profit / other ownership (NO/PRF)** exhibits a strong positive SHAP impact, suggesting higher survival likelihood relative to the baseline.

Interpretation: Legal structure influences resilience, likely due to differences in liability protection, financial resources, governance, and operational capacity.

3. Geographic variables (latitude, longitude, population density) show minimal impact

- Features num__lat and num__lng have very small SHAP spreads, clustered around zero.
- A slight variation exists within the county, and other variables capture the local context more effectively.

Interpretation: Raw spatial coordinates alone do not provide meaningful distinctions in survival outcomes. Geographic location matters in combination with other factors, such as population and business density.

4. Population and business density moderately influence survival

- Features num__population and num__business_density show measurable SHAP ranges.
- Higher population or business density may correspond to greater customer availability or competition pressures.

Interpretation: Local economic and demographic conditions have a less significant impact on survival compared to business-level characteristics, such as age and ownership type.

5. City name adds little predictive value

- City-level indicators (e.g., cat__address_city_SAN DIEGO) appear near the bottom of the importance ranking.
- SHAP values are tightly centered on the right of the zero with minimal positive directional influence.

Interpretation: After accounting for other variables (industry, business age, ownership type, population, and density), the specific city location does not significantly improve prediction accuracy. Differences within cities outweigh differences between cities.

6. NAICS sector meaningfully influences survival

Mid-ranking features include sectors such as:

- Transportation & Warehousing
- Health Care & Social Assistance
- Construction
- Retail Trade
- Professional, Scientific & Technical Services

Some sectors are driven by predictions of survival (e.g., healthcare, professional services), while others are driven by predictions of failure (e.g., construction, retail).

Interpretation: Industry-level risk is an important predictor, reflecting market stability, demand dynamics, and operational complexity.

7. Some features have a negligible impact

- Bottom-ranking features include rare ownership categories (NO/PRF, PARTNR), uncommon NAICS sectors, and most city-level indicators.
- SHAP magnitudes are near zero → minimal predictive influence.

Interpretation: These variables contribute little to predictive value due to their low frequency or weak discrimination.

8. Conclusion

The SHAP analysis indicates that business survival is primarily driven by **business-level characteristics**:

1. **Business age** – The most dominant factor.
2. **Ownership type** – Legal structure significantly influences resilience.
3. **Local economic context** – Population and business density matter moderately.
4. **Industry sector** – NAICS classification affects risk profiles.

Meanwhile, **geographic identifiers**, including latitude, longitude, and city name, contribute little once other factors are taken into account.

Overall, in this dataset, **organizational structure, business maturity, and industry context are far more significant for** survival than geographic location within San Diego County.

3.2 Unsupervised: Clustering (K-Prototypes, DBSCAN, K-Means)

Technique	Advantage	Disadvantage
Elbow Method	Simple and intuitive visualization. Helps determine the "optimal" number of clusters for K-Means or K-Prototypes. Less computationally expensive than the silhouette method, especially for large datasets.	Subjective. "elbow" is not always clear. Doesn't work well if clusters are not well-separated or vary in size/density.
K-Prototypes	Can handle mixed data types (numerical + categorical) (Huang, 1998). Scales to medium-large datasets.	Requires specifying the number of clusters k. Sensitive to initialization; may converge to local minima. Works poorly with very high-dimensional categorical data.
DBSCAN	Detects arbitrarily shaped clusters. Automatically identifies outliers. No need to specify the number of clusters.	Sensitive to **epsilon** and min **pts** parameters. Struggles with varying density clusters. Not ideal for high-dimensional data.
K-Means	Simple, fast, and widely used. Works well for well-separated spherical clusters.	Needs the number of clusters k in advance. Sensitive to outliers and initialization. Only works with numerical data; assumes roughly equal cluster sizes.

K-Prototypes:

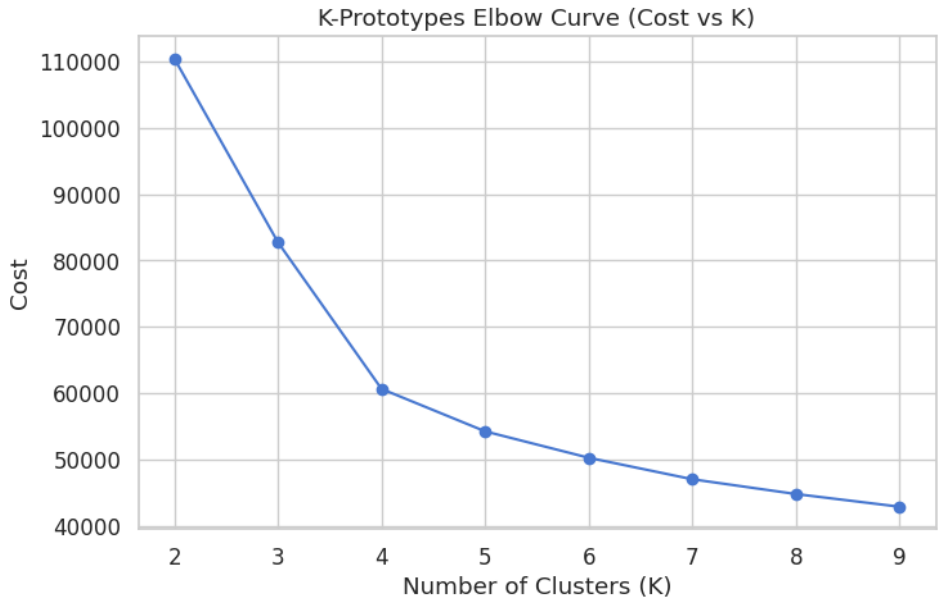
K-Prototypes is a clustering method designed for **mixed-type data**, meaning it can handle both **numerical variables** (e.g., business_age, population, business_density) and **categorical variables** (e.g., naics_sector_standard, ownership_type) simultaneously (Huang, 1998).

Unlike K-Means, which only works on numeric data, K-Prototypes calculates distances using a combination of:

- **Euclidean distance** for numeric features
- **Matching dissimilarity** for categorical features

This allows clusters to reflect patterns in both **quantitative metrics and categorical business types** (Huang, 1998).

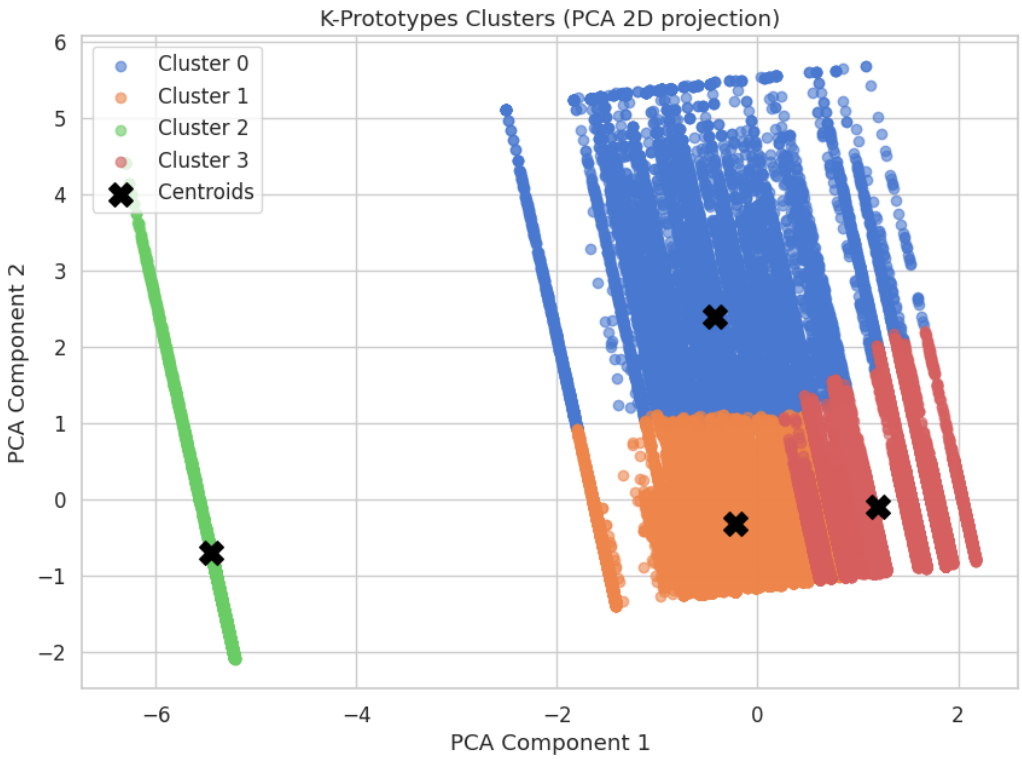
K-Prototype clustering was applied to capture patterns in both numeric (business_age, population, business_density) and categorical (NAICS sector, ownership type) features. Using the elbow method (see picture below), k=4 clusters were chosen.



The resulting clusters revealed distinct business profiles, such as older professional sole proprietorships in lower-density areas (Cluster 0) or high-density corporate hubs in professional services (Cluster 2). This method allowed us to account for mixed data types and generate interpretable business clusters. Here's what the clusters reveal:

Cluster	business_age	population	business_density	naics_sector_standard	ownership_type
0	2.42	-0.16	-0.12	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	SOLE
1	-0.28	-0.42	-0.04	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	SOLE
2	0.27	-2.24	5.52	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	CORP
3	-0.28	1.24	-0.41	RETAIL TRADE	SOLE

Plot for quantitative metrics from each cluster:



Interpretation of clusters

1. **Cluster 0:**
   - **Older businesses** (positive business_age)
   - Slightly **below-average population and density**
   - Predominantly **professional services**, owned as **sole proprietorships**

2. **Cluster 1:**
   - **Slightly younger businesses**
   - Lower-than-average population and density
   - Also, professional services, mostly **sole proprietorships**
   - Could represent **small-scale professional startups**

3. **Cluster 2:**
   - **Moderate-aged businesses**
   - Located in **low-population areas**
   - **Very high business density** → possibly industrial or corporate hubs
   - Most businesses are **corporations**

4. **Cluster 3:**
    - **Slightly younger businesses**
    - Found in **higher-population areas**
    - Lower business density
    - Belongs mainly to the **retail sector**
    - Mostly **sole proprietorships**

Why K-Prototypes is suitable

1. Mixed data types: The dataset combines numeric and categorical features. K-Prototypes can cluster all relevant information simultaneously (Huang, 1998).

2. Meaningful clusters: The results show distinct business profiles, not just based on numeric metrics but also on the type of business and ownership structure.

3. Interpretability: Each cluster can be described in plain language, which is valuable for business location analysis.

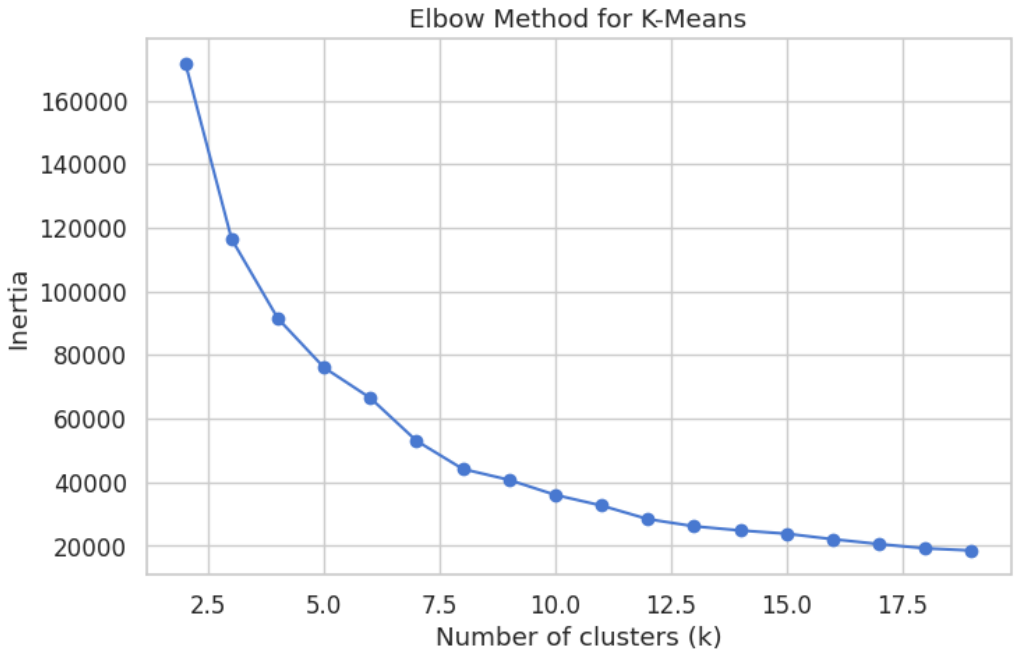
K-Means:

K-Means is a clustering method designed for **numeric data only**, meaning it works with continuous variables such as latitude, longitude, business age, population, or business density.

- It partitions the data into **k clusters** by minimizing the sum of squared distances between each point and the centroid of its assigned cluster.
- All calculations are based on **Euclidean distance**; therefore, categorical variables (such as NAICS sector or ownership type) cannot be used directly.
- **Categorical variables** (e.g., NAICS sector, ownership type) are summarized using the **mode** in each cluster.

Clusters produced by K-Means reflect **patterns in numeric features**, such as geographic concentration or business metrics. Although categorical variables did not influence cluster formation, they provide useful context regarding the characteristics of businesses within each geographic cluster.

K-Means clustering was applied to latitude and longitude to identify geographic concentrations of businesses. Using the elbow method, k=7 clusters were selected.



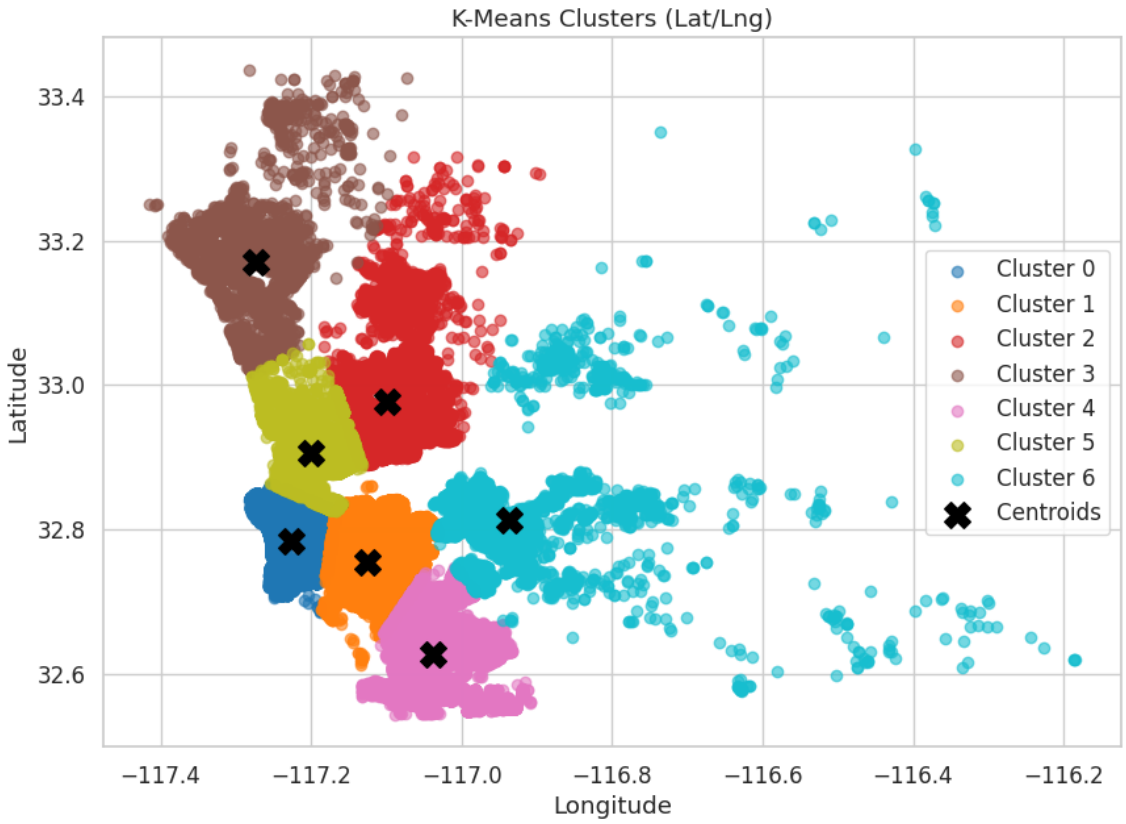
Although DBSCAN is well-suited for detecting arbitrarily shaped clusters and filtering out noise, it did not perform effectively on this dataset. When applied to the spatial features (latitude and longitude), DBSCAN produced **133 clusters**, largely because the density of businesses varies greatly across San Diego (dense urban cores vs. sparse suburban and rural areas). The algorithm is highly sensitive to the *eps* parameter; small adjustments drastically change the number of clusters, and no stable clustering structure emerges. Because of this instability and the large number of clusters generated, DBSCAN did not provide interpretable or meaningful regional groupings. Therefore, **K-Means was**

selected instead, as it produced a manageable number of clusters that aligned with the geographic patterns in the data and provided clearer, more consistent segmentation suitable for analysis and presentation. Here's what the k-means clusters reveal:

Cluster	num_businesses	avg_lat	avg_lng	avg_business_age	avg_population	avg_business_density	naics_sector_standard_mode	ownership_type_mode
0	23,754	32.784	-117.226	9.70	38,446	114.86	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	SOLE
1	55,919	32.754	-117.125	8.80	43,015	119.41	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	SOLE
2	17,362	32.978	-117.099	9.30	52,985	61.91	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	SOLE
3	2,454	33.171	-117.273	8.39	40,636	5.36	CONSTRUCTION	SOLE
4	15,850	32.627	-117.037	7.74	70,388	52.18	RETAIL TRADE	SOLE
5	17,391	32.906	-117.201	9.76	45,875	227.83	PROFESSIONAL, SCIENTIFIC, AND TECHNICAL SERVICES	SOLE

6	5,084	32.813	-116.937	8.66	45,744	22.60	CONSTRUCTION	SOLE

K-means Clusters Plot :



The K-Means clustering algorithm was applied to the latitude and longitude coordinates of all businesses to identify spatial patterns across San Diego County. After computing the cluster centroids, regional labels (e.g., North County, Central, East County, South Bay) were assigned qualitatively based on the average latitude and longitude of each cluster. These labels are used purely for interpretive purposes to help contextualize the spatial distribution of businesses. They do not represent official administrative boundaries and are not derived from TIGER/US Census regional shapefiles. Instead, they reflect broad geographic tendencies commonly used in public, academic, and local discussions of San Diego's layout.

Geographic interpretation of the clusters is based on the natural ordering of latitude and longitude. In the San Diego region, **latitude increases as you move north** and decreases as you move south, while **longitude becomes less negative as you move east** and more negative as you move west. Using these spatial conventions, each cluster's centroid (its average latitude and longitude) was compared to the county's geographic layout to infer its approximate regional position. For example, clusters with higher average latitudes correspond to areas in North County, while clusters with lower latitudes tend to fall toward the South Bay. Similarly, clusters with longitudes closer to –116.9 lie toward East County, whereas those around –117.25 represent coastal or western areas. Applying these rules to the K-Means centroids produced the following regional interpretations:

Cluster	avg_lat	avg_lng	Interpretation
0	32.784	-117.226	Central-West
1	32.754	-117.125	Central-East
2	32.978	-117.099	North-East
3	33.171	-117.273	North-West
4	32.627	-117.037	South-East
5	32.906	-117.201	North-Central-West
6	32.813	-116.937	Central-East / East County Edge

Interpretation of clusters:

Cluster 0

- Moderate-sized cluster **(23,754 businesses).**
- Located around **32.78° N, –117.23° W** → *Central-West*.
- Average business age**: ~9.7 years** (older businesses).
- Average population: **~38k.**
- Business density: **114.86 (high).**
- Predominantly **professional, scientific, and technical services;** mostly **sole proprietorships.**

Cluster 1

- Largest cluster **(55,919 businesses).**
- Located around **32.75° N, –117.13° W** → *Central-East*.
- Average business age: **~8.8 years.**
- Average population: **~43k.**
- **Business density: 119.41** (slightly higher than Cluster 0).
- Mostly **professional, scientific, and technical services;** dominated by **sole proprietorships.**

Cluster 2

- Medium-sized cluster **(17,362 businesses).**
- Located around **32.98° N, –117.10° W** → *North-East*.
- Average business age: **~9.3 years (older businesses).**
- Average population: **~53k.**
- Business density: **61.91 (moderate).**
- Primarily **professional, scientific, and technical services;** with many **sole proprietorships.**

Cluster 3

- Smallest cluster **(2,454 businesses).**
- Located around **33.17° N, –117.27° W** → *North-West*.
- Average business age: **~8.4 years.**
- Average population: **~41k.**
- Business density: **5.36** (very low; dispersed area).
- Primarily the **construction sector;** mostly **sole proprietorships.**

Cluster 4

- Medium-sized cluster **(15,850 businesses).**
- Located around **32.63° N, –117.04° W** → *South-East*.
- Average business age: **~7.7 years** (younger businesses).
- Average population: **~70k** (highest population among clusters).
- Business density: **52.18** (moderate).
- Dominated by **retail trade,** mostly **sole proprietorships.**

Cluster 5

- Medium-sized cluster **(17,391 businesses).**
- Located around **32.91° N, –117.20° W** → *North-Central-West*.
- Average business age: **~9.8 years** (oldest cluster overall).
- Average population: **~46k.**
- Business density: **227.83** (highest density of all clusters).
- Primarily **professional, scientific, and technical services;** mostly **sole proprietorships.**

Cluster 6

- Small-sized cluster **(5,084 businesses).**
- Located around **32.81° N, –116.94° W** → *Central-East / East County edge*.
- Average business age: **~8.7 years.**
- Average population: **~46k.**
- Business density: **22.60** (low).
- Mainly **the construction sector,** dominated by **sole proprietorships.**

Why K-Means is suitable

1. **Spatial focus:** K-Means identifies **geographic clusters**, enabling the analysis of how businesses are concentrated in space.

2. **Simpler than DBSCAN:** In the data, DBSCAN created **too many clusters (133)** due to sensitivity to eps. K-Means provided **7 interpretable clusters**.

3. **Cluster interpretability:** By summarizing numeric averages (business age, population, density) and categorical modes (NAICS sector, ownership), each cluster can be described clearly for business analysis.

4. **Scalable:** Works efficiently with large datasets.

DBSCAN **clusters (133)**

Data Summary and Implications

The data analysis revealed key insights into business survival, activity, and spatial distribution in San Diego County. By combining categorical and numerical analyses with supervised and unsupervised machine learning models, the study provided a comprehensive view of factors influencing business outcomes.

Implications for Research Question:
Logistic regression was initially proposed as the predictive modeling approach due to its interpretability. However, the logistic regression model achieved an accuracy of 66.2%, which is below the 70% threshold set in the alternative hypothesis (H_1). Therefore, the null hypothesis (H_0) that a predictive logistic regression model cannot be constructed **cannot be rejected**. This indicates that logistic regression alone is insufficient for reliably predicting business survival in this dataset.

Tree-based ensemble methods, particularly LightGBM, were also evaluated and achieved substantially higher predictive performance (accuracy = 0.836, Precision = 0.828, Recall = 0.760, F1 Score = 0.793, ROC-AUC = 0.911). These results demonstrate that nonlinear, tree-based models are more effective at capturing complex relationships between business characteristics, geographic features, and survival outcomes.

The strong performance of LightGBM suggests that predictions from this model can inform practical business applications, such as identifying high-risk businesses that may need support and guiding strategic expansion decisions. By leveraging insights into key predictors such as business age, ownership type, industry sector, and local economic conditions, entrepreneurs and investors can more effectively allocate resources, select optimal locations for expansion, and mitigate the risk of closure.

Key Findings:

Business Survival Drivers:
SHAP analysis from the LightGBM model identified business age as the most critical predictor of survival. Older businesses consistently demonstrated higher survival rates, while younger businesses exhibited higher closure risk. Ownership type also significantly influenced survival, with corporations (Corp/S Corp), NO/PRF (but with low impact), and LLCs generally showing greater resilience than Sole Proprietorships, NO/PRF, H-W, and PARNTR proprietorships. The industry sector had a moderate impact; Health Care and Social Assistance were more stable than Transportation or Professional Services. Geographic variables, including latitude, longitude, and city, showed minimal predictive value, suggesting that **survival is primarily driven by internal business characteristics rather than location alone.**

Spatial Business Clusters:
K-Means clustering of latitude and longitude revealed seven distinct geographic clusters, capturing concentrations of businesses across San Diego. In contrast, DBSCAN produced 710 clusters due to varying densities, which were not interpretable. K-Means allowed for clear regional groupings (e.g., North County, Central, East County, South Bay) and highlighted areas of high business density (e.g., North-Central-West cluster) versus low-density, dispersed regions (e.g., North-West cluster).

Business Profiles from K-Prototypes:
K-Prototypes clustering, which accounts for both numeric and categorical features, identified four meaningful business clusters. These clusters reflected patterns such as older professional services in lower-density areas versus corporate hubs with high business density. This confirmed the heterogeneity of business types across demographic and operational characteristics.

Limitation:
One limitation of this analysis is the use of K-Means clustering, which is based solely on latitude and longitude, to infer regional groupings. Approximate labels (e.g., North County, East County) were inferred from cluster centroids and may not perfectly align with official TIGER/US Census boundaries, potentially affecting precise geographic targeting or policy implementation.

Recommended Course of Action
Based on the analysis, business development initiatives should focus on supporting younger and sole proprietorship businesses, especially in lower-density or higher-risk areas. Programs could include:

- **Mentorship and Advisory Services:** Provide guidance on operational best practices, financial management, and strategic planning to improve survival rates.
- **Access to Capital and Incentives:** Facilitate funding opportunities, grants, or low-interest loans to help small and new businesses scale operations and invest in growth.
- **Sector-Specific Support:** Offer tailored training, market analysis, and resources for industries identified as higher risk, such as retail and construction, to strengthen resilience and reduce the probability of closure.
- **Data-Informed Expansion Facilitation:** Utilize insights from spatial clustering and demographic data to identify high-density business clusters or underserved markets where expansion is likely to be successful. This enables entrepreneurs and investors to make informed, targeted decisions for safe and strategic business growth.

Expected Benefits of the Study

This study offers clear, practical benefits for **entrepreneurs and investors** by identifying the most influential factors driving business survival in San Diego County and demonstrating how predictive analytics can support informed decision-making.

The results show that survival is primarily driven by internal business characteristics, especially **business age**, **ownership type**, and **industry sector,** rather than geographic location alone. This insight enables entrepreneurs to more accurately assess operational risk and allows investors to evaluate business viability more effectively.

The superior performance of the LightGBM model, supported by SHAP analysis, provides a reliable and interpretable tool for **predicting business survival in practice**. Entrepreneurs can use model outputs to identify early-stage or high-risk ventures that may require strategic adjustments, mentorship, or capital infusion, while investors can leverage these predictions to prioritize funding toward more resilient business profiles and mitigate closure risk.

Additionally, spatial and clustering analyses offer strategic value by highlighting business density patterns, regional groupings, and distinct business profiles across the county. These insights help entrepreneurs identify favorable markets for expansion and assist investors in recognizing high-potential clusters and underserved areas. Overall, the study delivers a data-driven framework that supports smarter investment decisions, targeted business support, and sustainable entrepreneurial growth.

Directions for Future Study:

1. **Business Location and Market Potential Analysis:** Extend the dataset to include additional geographic and economic context, such as proximity to competitors, foot traffic, commercial zoning, or customer demographics. By combining business survival data with local market characteristics, future studies could identify optimal locations for business expansion, assess market saturation, and target areas with the highest growth potential. This approach moves beyond administrative boundaries and directly informs strategic expansion and resource allocation.

2. **Temporal and Trend Analysis:** Given that the dataset spans from 1974, it contains longitudinal information that can be leveraged for time-based analysis of business activity and survival. By structuring the data by year (or quarter), trends in business openings, closures, and survival rates can be examined over time. Key temporal metrics, such as the annual number of active businesses, closure rates, average business age, and business density, can be computed for each region, NAICS sector, or ownership type.

Time-series analysis of these metrics can reveal patterns such as emerging growth areas, periods of elevated closure risk, or sector-specific trends influenced by economic cycles. Additionally, incorporating temporal features into predictive models, such as LightGBM, could improve the estimation of survival probabilities and help identify the most promising periods and locations for business expansion.

Visualization of these trends would further support strategic decision-making by highlighting areas with sustained growth or persistent risk, enabling entrepreneurs and investors to make data-driven decisions on resource allocation and design targeted business support or expansion strategies.

References

- City of San Diego. (n.d.). *San Diego business listings dataset*. City of San Diego Open Data Portal. Retrieved November 3, 2025, from https://data.sandiego.gov/datasets/business-listings
- Corporate Finance Institute. (2023). *Business lifecycle stages and analysis*. Retrieved November 3, 2025, from https://corporatefinanceinstitute.com/resources/valuation/business-life-cycle/
- García-Vidal, G., Sánchez-Rodríguez, A., Guzmán-Vilar, L., Pérez-Campdesuñer, R., & Martínez-Vivar, R. (2025). Toward building a model of business closure intention in SMEs: Binomial logistic regression. *Administrative Sciences, 15*(7), 240. Retrieved November 4, 2025, from https://doi.org/10.3390/admsci15070240
- Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283–304. Retrieved November 4, 2025, from https://doi.org/10.1023/A:1009769707641
- IBM. (n.d.). *Python vs. R for Data Science: A comprehensive comparison*. IBM Think. Retrieved November 3, 2025, from https://www.ibm.com/think/topics/python-vs-r
- Nemade, B., Bharadi, V., Alegavi, S. S., & Marakarkandy, B. (2023). A comprehensive review: SMOTE-based oversampling methods for imbalanced classification techniques, evaluation, and result comparisons. *International Journal of Intelligent Systems and Applications in Engineering, 11*(9 s), 790–803. Retrieved November 4, 2025, from https://ijisae.org/index.php/IJISAE/article/view/3268
- Querio.ai. (2025, September 7). *Statistical tools for data analysis: When to use R, Python, SPSS, or SAS*. Business Intelligence. Retrieved November 4, 2025, from https://querio.ai/articles/statistical-tools-for-data-analysis-when-to-use-r-python-spss-or-sas
- Ramirez, K. (2025). *San Diego County Active/Closed Business Listing* [Data set]. Kaggle. Retrieved November 12, 2025, from https://www.kaggle.com/datasets/kenierramirez/san-diego-county-activeclosed-business-listing
- Reyes, M., & Suárez, L. (2022). Logistic regression analysis of firm survival in Mexico City. *Journal of Business Analytics, 5*(2), 55–71. Retrieved November 4, 2025, from https://www.sciencedirect.com/science/article/pii/S2405844022005072
- SANDAG. (n.d.). *2020 Census population by ZIP code dataset*. San Diego Association of Governments Open Data Portal. Retrieved November 3, 2025, from https://opendata.sandag.org/Census/2020-Census-Population-by-ZIP-Code/26f5-2x9a/about_data

- Society of Actuaries. (2019). *Considerations in predictive modeling: Feature engineering and selection* (Research Report). Society of Actuaries. Retrieved November 5, 2025, from https://www.soa.org/globalassets/assets/files/resources/research-report/2019/considerations-predictive-modeling.pdf

- U.S. Census Bureau. (2024, February 12). *NAICS codes & understanding industry classification systems*. Retrieved November 4, 2025, from https://www.census.gov/programs-surveys/economic-census/year/2022/guidance/understanding-naics.html