Refer to Appendix-1 for Abbreviation.

1. Question-1

1.1 Data-preparation

CRISP-DM's 6-step approach is widely-adopted as industry-standard, amidst evoluting analysis-models since 1990s [1], [2], [3], being guidelines of this project based on Chapman et al.'s original-proposal [4].

Within context, this section explores business-logic and data-understanding stages, following scenarios as property-agency and rising-rent environment [5], therefore exploring correlation between rent and key-influential factors. Given that raw-and-unstructured datasets required for cleaning before machine-learning implementation: (1) mistakenly-labelled test-dataset columns, unfitted for model-alignment [6]; (2) duplicated/missing/inconsistent data, enlarging bias with weakened model-performance; (3) fragmented formats, limiting interpretation and insights [7]; and (4) outliers, impairing model-precision with bias [8].

Therefore, data-cleaning standardise formats and rectify errors for accuracy /consistency [9]. Referring to Hellerstein's data-cleaning guidelines and Microsoft-Support's checklist to avoid missing-procedures [10], [11]. Feature-engineering improves data-analysis through grouping and re-calculating factors; and splitting full-addresses into street-based/town-based categories [12], [13]. Noise-reduction is exercised with KNN-imputing and outliers-scaling with IQR and z-score, boosting model-performance with less deviation *(see Figure-1-3)* [14]. Consequently, valid-records are scaled to 16196 for subsequent-analysis *(Appendix-2)*

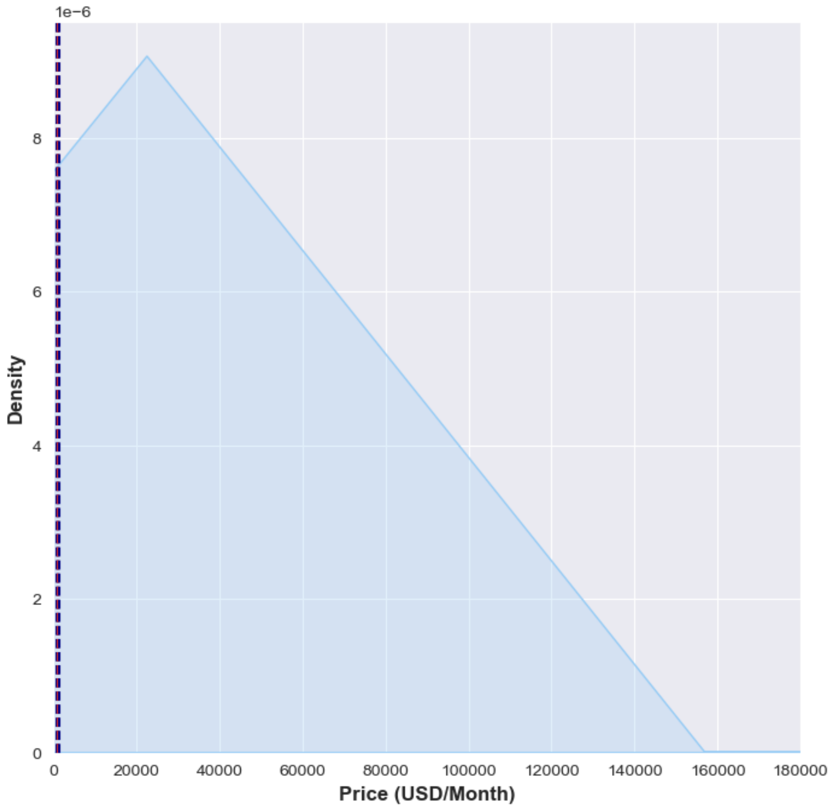

Figure-1: KDE plot: Distribution after Data-cleaning

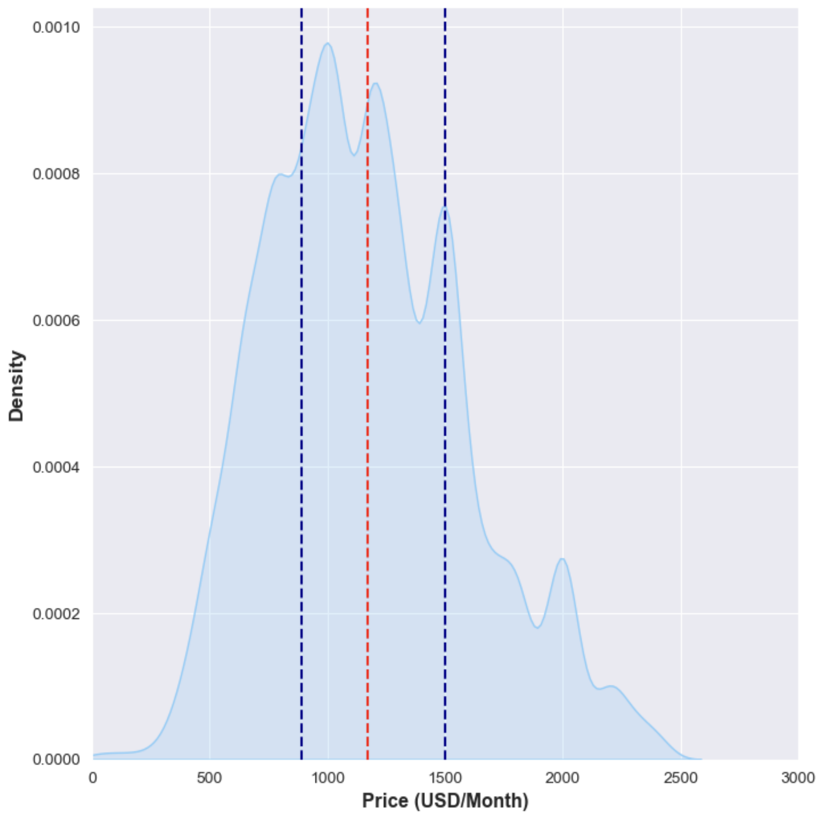

Figure-2: KDE plot: Distribution after Data-filtering

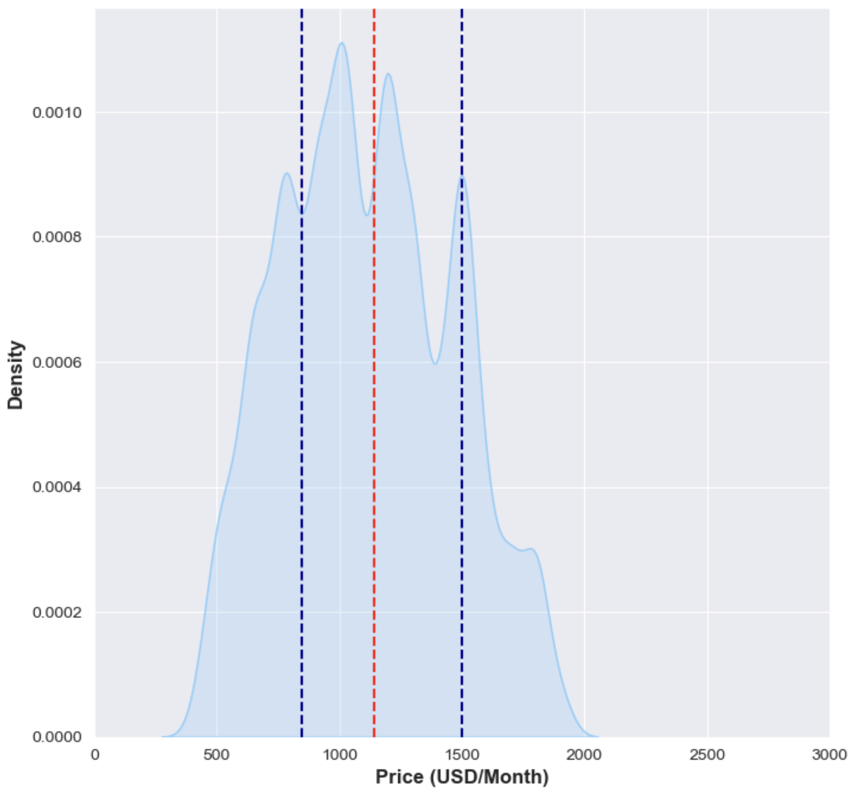

Figure-3: KDE plot: Distribution after Data-processing

Regarding modelling, RFC is chosen for machine-learning process. As *task-1(a)* and *task-1(b)* analyse discrete values, which classification-model can identify optimal categories-boundaries directly [15], [16]. Though regression-model matches *task-1(c)'*s continuous nature, lacking linear-pattern leave challenges due to restricted data-distribution. *Figure-1-3* illustrates that data-preprocessing effectively scales datasets and manages outliers; however, *Figure-4* indicates poor regressor-performance. Even best-performed GBR, only reaches R^2 at 0.41 - nearly halved academic-standard at 0.80, projecting highly-variability [17], [18].

	MAE	MSE	RMSE	R2 Score
Gradient Boosting Regressor	0.2989	0.1430	0.1430	0.4124
SVR Regressor	0.2715	0.1444	0.1444	0.4067
Random Forest Regressor	0.3166	0.1595	0.1595	0.3446
KNN Regressor	0.2411	0.2411	0.2411	0.0094

Figure-4: Table - Comparison of Classification Models

	Accuracy (%)	Precision (%)	Recall (%)	F1 index (%)
Random-forest Classifier	78.72	78.65	78.72	78.68
SVC Classifier	78.72	78.63	78.72	78.64
KNN Classifier	75.89	75.74	75.89	75.73
Decision-tree Classifier	74.58	75.00	74.58	74.70

Figure-5: Table - Comparison of Classification Models

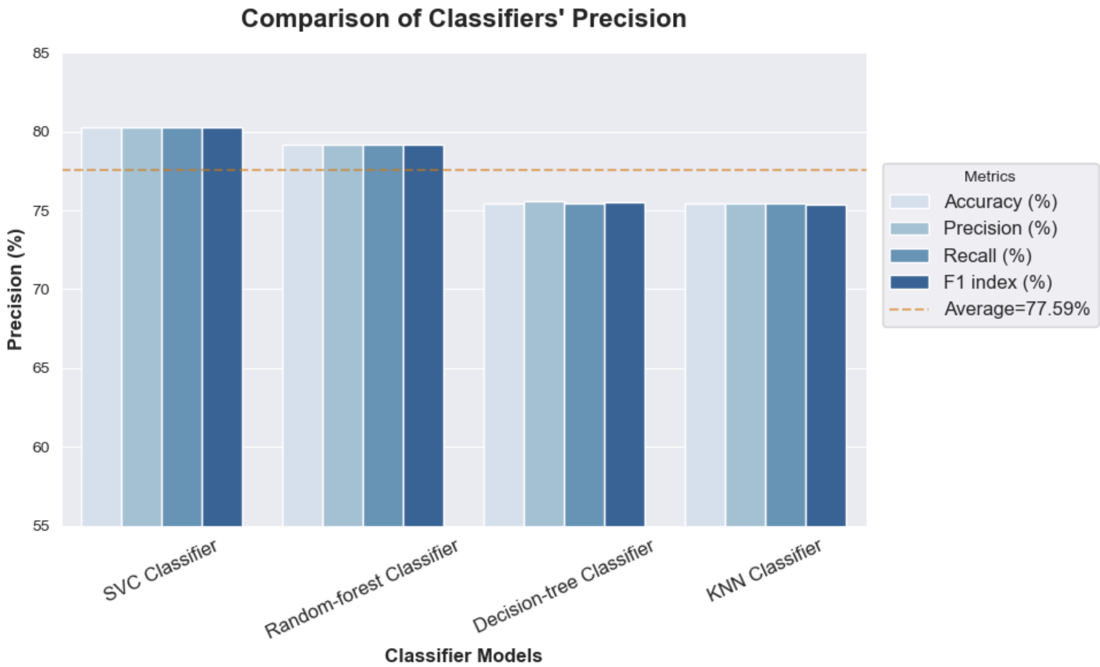

Figure-6: Barchart - Comparison of Classifier's Precision Matrix

Nevertheless, classifiers demonstrate highly-reliability by achieving approximately 75% in accuracy, precision, recall, and F1 *(Figure-5-6).* It satisfies *task-1(c)*'s criteria on analysing correlation to continuous-data, instead of trend-prediction.

RFC is selected for modelling among classifiers, approximate to SVC with stable-performance nearly 80% and diverse/dispersed data adopted *(Figure-5).* Provided that good-precision and recall, RFC can identify nearly 80% of samples, while producing output in similar accuracy [19]. Compared to SVC, RFC tolerates varied data-nature, especially involving discrete-categories/boundary-based data [20]. Built-in feature-importance further helps identifying key-influential factors, exploring correlations towards varying-rents as required [21]. Performance-stability, data-compatibility and interpretation-potentials make RFC best-suited ahead of other options.

1.2 Data-Analysis

1.2.1 Task-1(a)

Tasks-1(a) examines three discrete-factors correlating to rental-price. Filtering feature-factors, RFC adopted MDI and calculates feature-importance by averaging reducing impurity, analysing nonlinear-patterns and factors' interrelationships [22]. However, averaging continuous/categorical-variables may cause numerical-bias towards distortion [23].

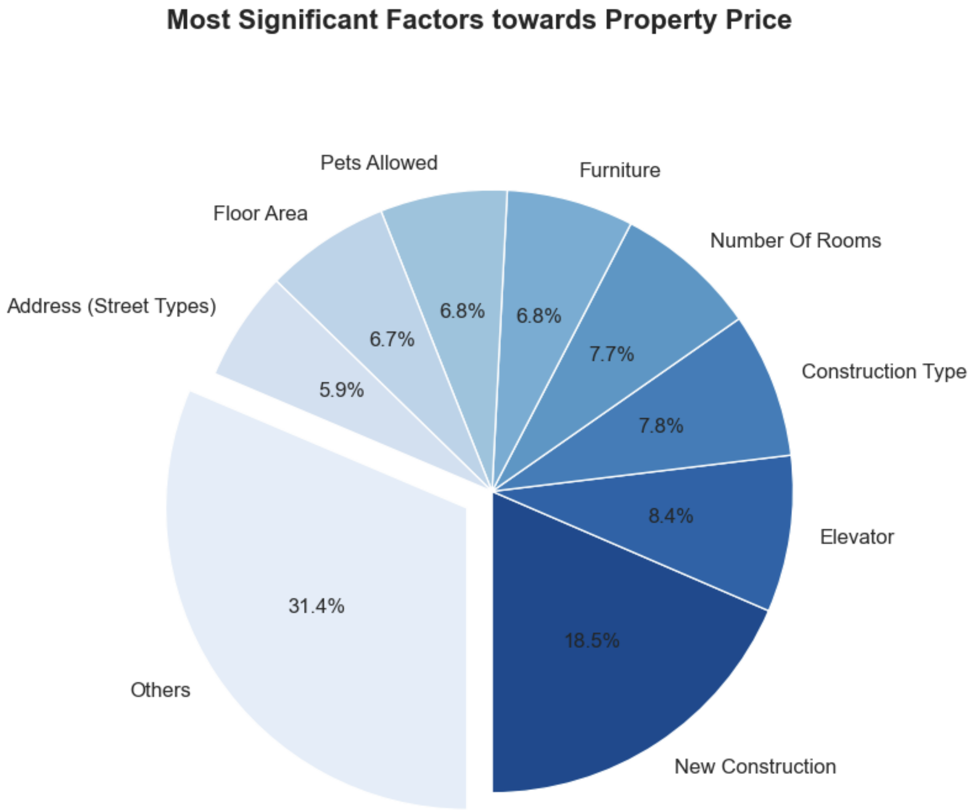

Figure-7: Piechart - Most Significant Factors towards Property Price

Accordingly, *Figure-7* filters top-eight influential-features upon feature-importances. *"New Construction"* accounts for nearly one-fifth of total-rate, revealing significant-demand for new-built properties. *"Elevator"* and *"Construction Types"* follow 8.4% and 7.8%, influencing rental-prices. These discrete/binary/categorical variables are used in analysis. Additionally, *"Others"* shared one-third of total importance, projecting notable and accumulative-impacts. Overall, major-features except top-three ranged between nearly 6%-9%, indicating rental-prices are influenced broadly with no single factor decided.

To explore data-distribution, boxplots well-fit for comparing rental-price and discrete-factors. Median, quartiles, ranges and outliers indicate price-variation, outlining existing patterns and highlighting price-dispersion across categories, identifying variance in trend/distributions [24]. However, weakness in explaining data-volume restricts its interpretations over big-data [25].

Figure-8: Boxplot - Correlation between Price and New Construction

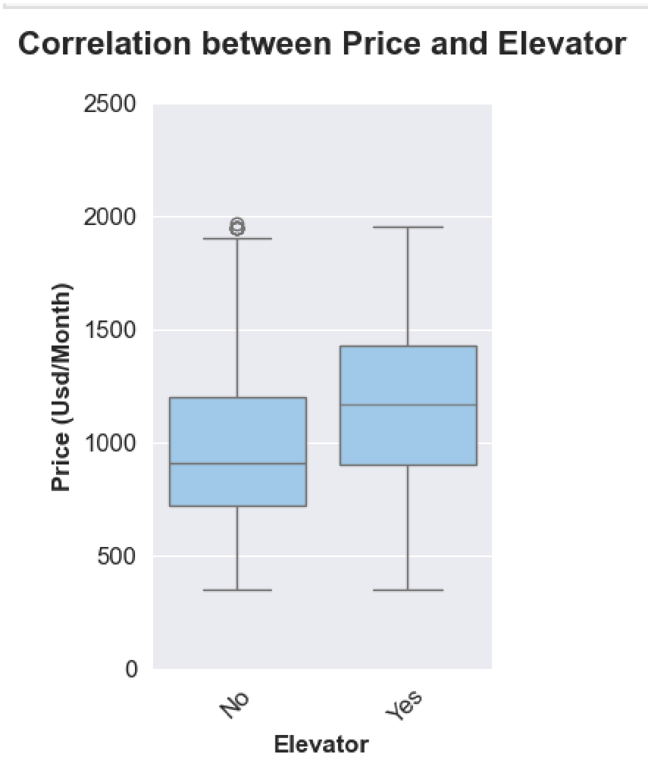

Figure-9: Boxplot - Correlation between Price and Elevator

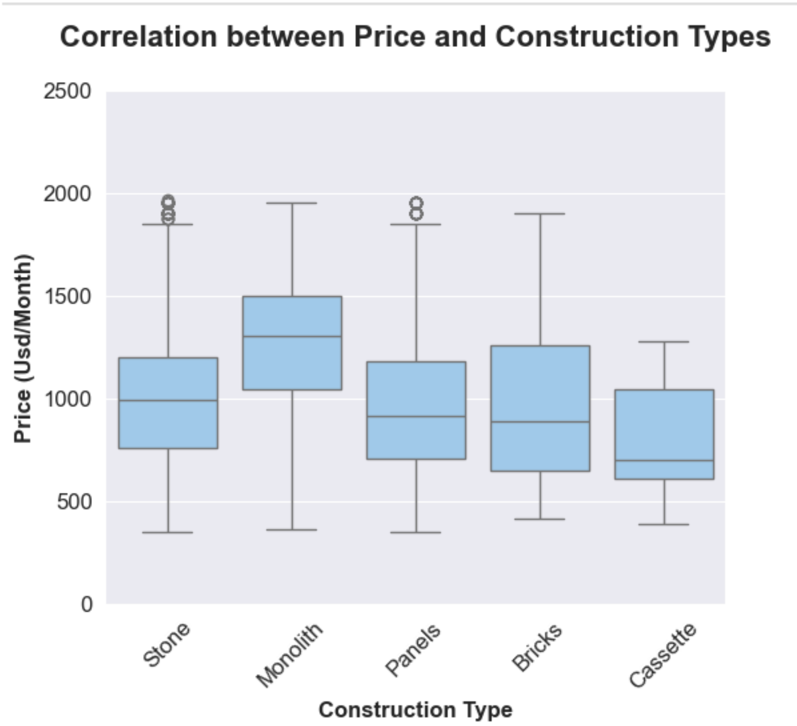

Figure-10: Boxplot - Correlation between Price and Construction Types

In *Figure-8-10*, selected discrete-variables ranged at \$500-\$2000 which is possibly caused by scaling/nominalisation. Whiskers imply how extreme-outliers and other variables' influences spread [26]. Specifically, rental-prices in *"New Construction"* and *"Elevator"* are approximately \$500 higher in positive-records than negatives. In *"Construction Types"*, Monolith has median nearly \$1300 with lower-quartile above other's median; conversely, Cassette has tight-ranged lower-quartiles below \$750, representing lower-segment focus.

The result suggests new-built properties, elevators and specific-materials have higher and concentrated rental-price range, revealing positive-correlation between rent and these influential-factors.

1.2.2 Task-1(b)

Task-1(b) examines inter-relationship between *"Price"*, *"Number of Rooms"* and *"Duration"*, through analysing monotonic-relationship with Spearman's-*"p"* and p-values statistic-significance. Unlike extreme-outliers-sensitive Pearson's-*"r"*, p-value ranking data-values is more adaptable with noise-data [27]. It offers stronger computing-efficiency than potential-alternatives like Kendall's-*"τ"* in big-data-processing [28]. However, overlapping-ties could distort data which is better cross-checking with additional-methods [Ibid., pp.84].

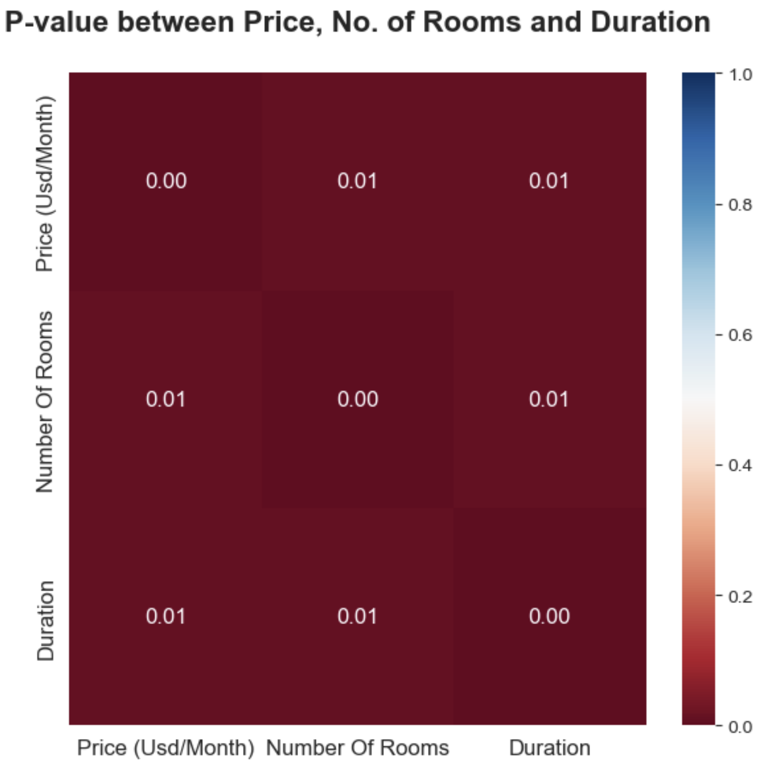

Figure-11: Heatmap - P-value between Price, Number of Rooms and Duration

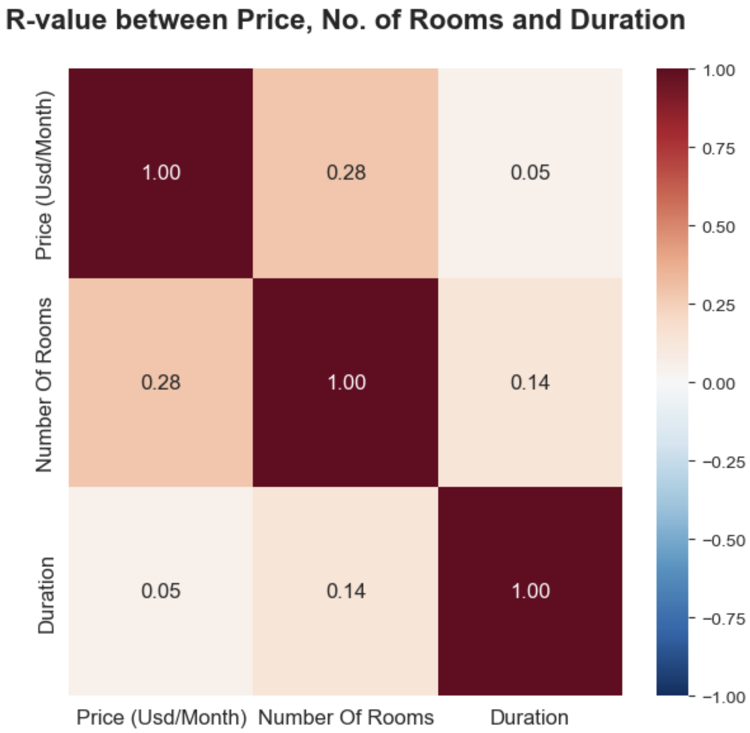

Figure-12: Heatmap - R-value between Price, Number of Rooms and Duration

Heatmap highlights relationships by using hue *(positive/negative)* and colour intensity *(influential-extent)*, while labelling coefficient-rates for precise-interpretations [29]. *Figure-11* showed, p-values below 0.05 standard highlights statistical-significance, rejecting null-hypothesis among these factors [30]. *Figure-12* further illustrates positive-correlation between selected-variables, while the pair *"Numbers of Room"* and *"Price"* reached 0.28 as weak-correlation with subtle-influences. Other pairs show r-value nearly 0, representing negligible correlation and influences. This finding aligns to Figure-7 that "Number of Rooms" ranked fourth *(7.7%)* in feature-importance, while *"Duration"* is insignificant. This evidence well-supported explanation that the former have higher-relevance towards rental-prices.

Boxplot well-performs in analysing multiple-variable-types distribution. It is chosen rather than violin-plots, which explains variation with median, interquartile and outliers [26]. Similar to Spearman's-*"p"*, boxplots are less-sensitive to outliers and not depend on averaging-data; visualising real-values to implement rank-patterns for better interpretations [31], [32].

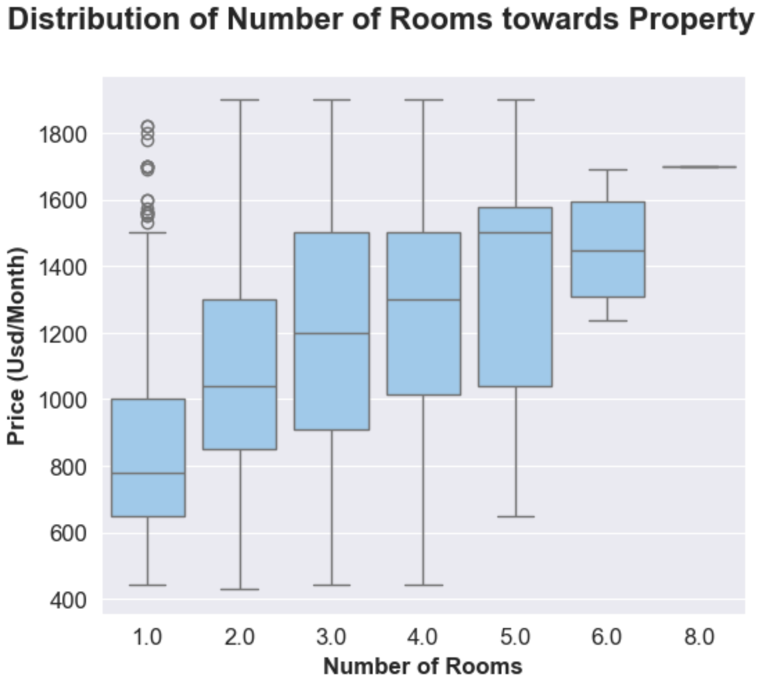

Figure-13: Boxplot - Distribution of Number of Rooms towards Property Price

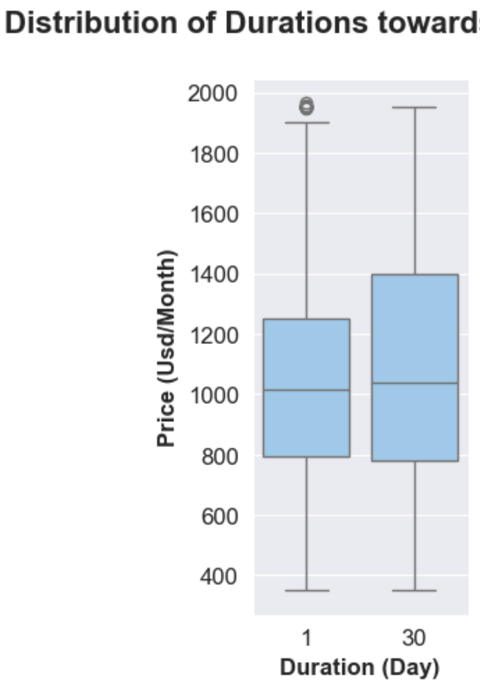

Figure-14: Boxplot - Distribution of Duration towards Property Price

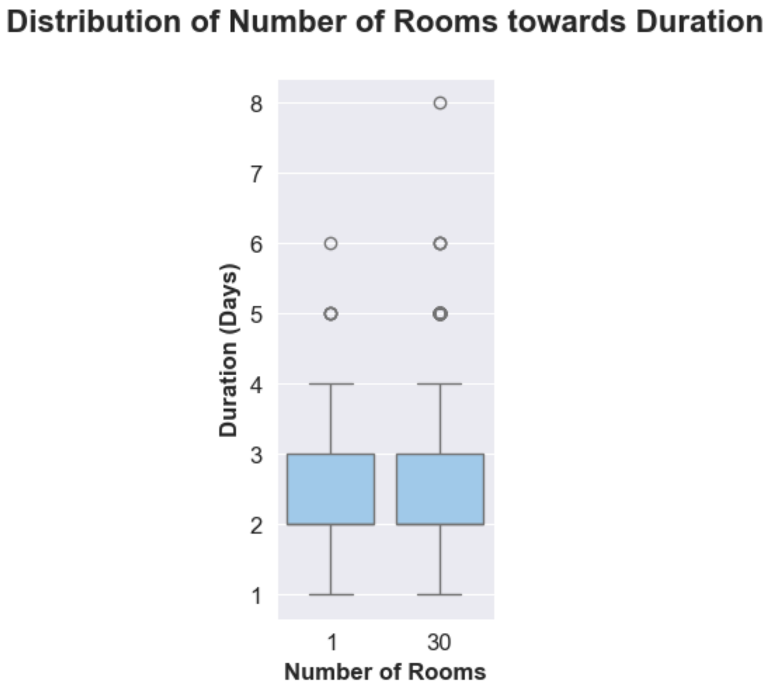

Figure-15: Boxplot - Distribution of Number of Rooms towards Duration

Figure-13 shows rising rent-price with range 1-4 rooms, as median boosted nearly one-third to approximately $1300 and lower-quartile rise correspondingly. It reveals general price-rise in rental-market. However, weakened-correlations existed due to smaller differences beyond 5-rooms with volatility. *Figures-14-15* reveal similarly in weak-correlation, as monthly-rent slightly-grows in upper-quartile and outliers of room-counts flavoured monthly-rental, potentially-linked to these variables.

1.2.3 Task-1(c)

Task-1(c) examines whether address is key-influential variable to rental-price, alongside-with previously-studied factors. ANOVA was adopted to assess the significant-differences of average rental-price across selected feature-factors [33]. By comparing between-group and within-group variance with f-statistic and p-values, this method prevents Type-I error inflation that caused by multiple t-tests and evaluates multiple factors simultaneously [34].

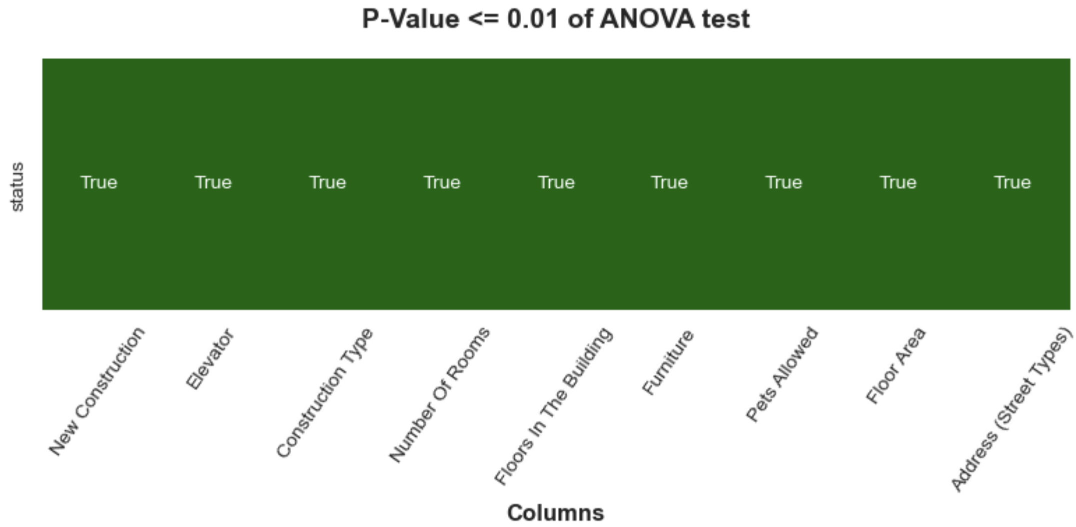

Figure-16: Heatmap - P-value Less Than 0.01 in AVOVA Analysis

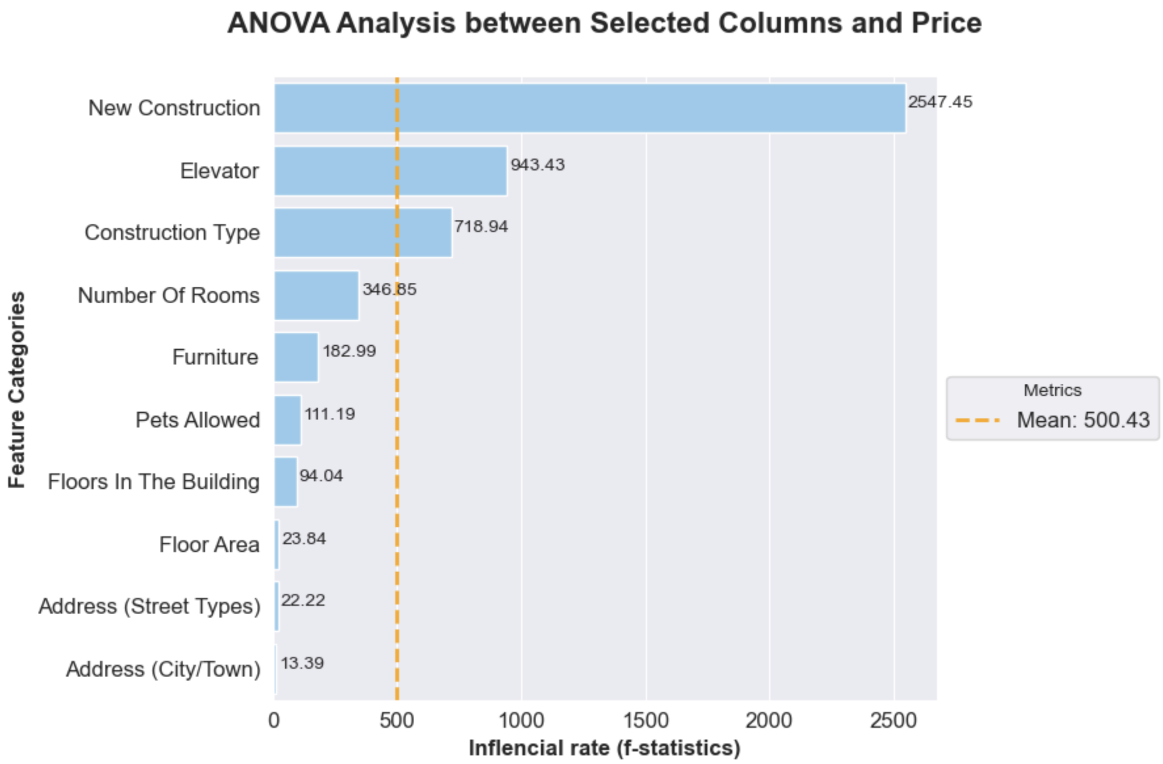

Figure-17: Barchart - F-statistics between Selected Columns and Price

The analysis adopted one-way ANOVA for testing *"Address (Street Types)"* with features in *task-1(a).* In *Figure-16*, p-values below 0.01 ensure statistical-significances of variables, as steps done in *task-1(b)* [30]. Figure-17 highlights *"New Construction"* as most-contributed factors with f-statistics over 2500 - over double than second-place. *"Elevator"* and *"Construction Type"* also showed strong effects over the average index (500.43). However, *"Address (Street Types)"* ranked lowest at around 20, aligning with low feature-importance in *Figure-7 (insignificant-part in "Others")*.

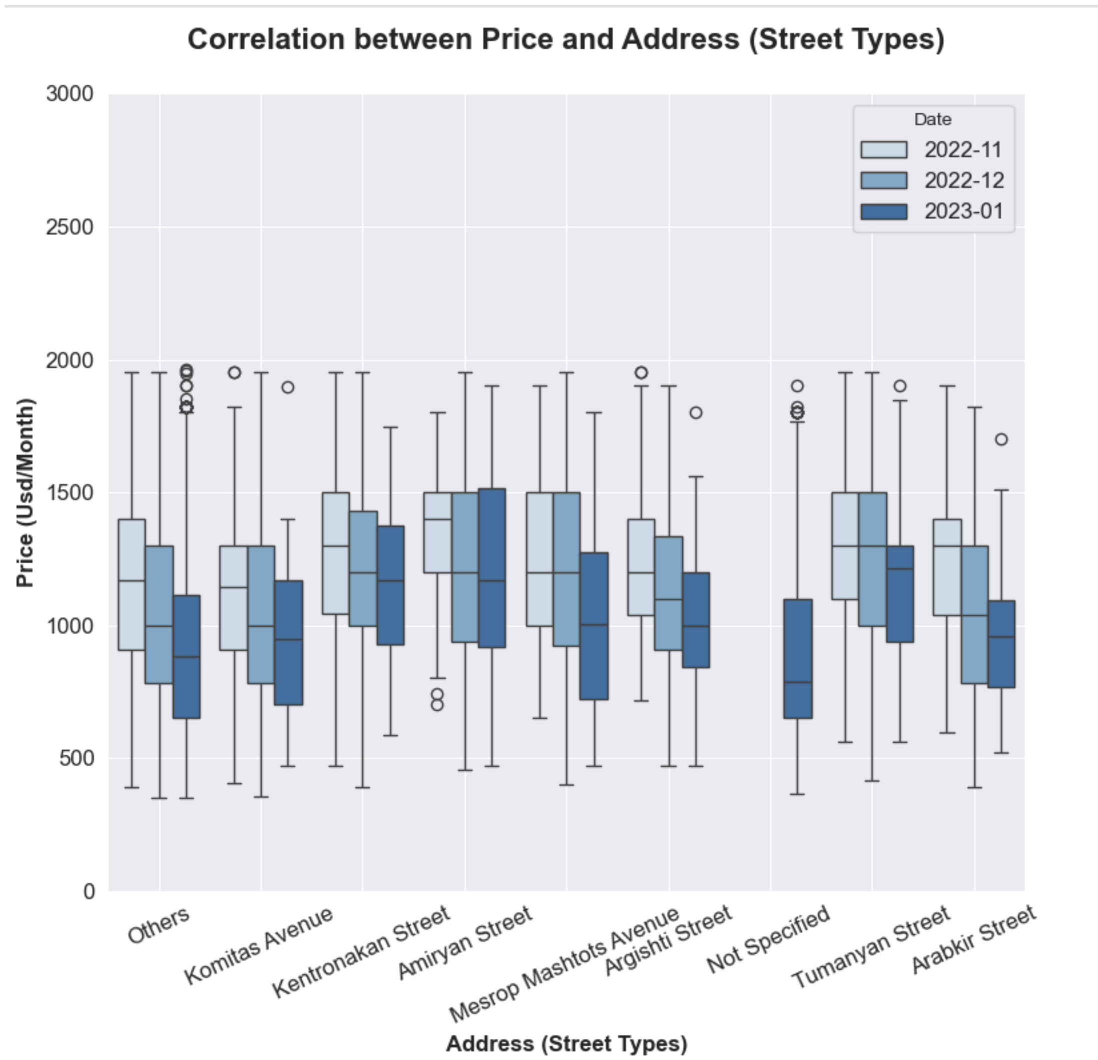

Figure-18: Boxplot - Correlation between Price and Address (Street Types)

In *Figure-18*, boxplot further shows variation of real-data, assessing whether address is negligible in explaining price-variation. Addresses with highest-counts ranged mostly between around $700-$1500, projecting weak-correlation and variation towards rental-prices. The finding indicates influences of street-names are restricted towards varied-prices. By splitting months, *Figure-18* further shows no distinct-tendency is found with featured-addresses, and most of them thereby sliding by month and generally lowest in “2023-01”. No proportion among these categories can be clearly found. Weak-relationship between addresses likely caused by variable’s complexity and imbalance data-sizes.

Unlike the top-three features studied in *task-1(a)*, analysis on *addresses* does not show proportional-relationship among its subcategories, as our observation in *Figures-8-10*. *“New Construction,” “Elevator,”* and *“Construction Type”* showed strong-correlations with varied-prices, Weak and inconsistent-patterns in address reveals its prediction is restricted, which supported earlier findings that addresses is the leastinfluential factors *(among selected-features)* towards rental pricing and concluded address is not significant-predictor of rent-inflation.

1.3 Evaluation

This project contains significant data-overcleaning before machine-learning process. Non-conforming records were strictly-deleted for completeness of crucial-fields *(Price and Address)*, slashed from 32587 to below 16196 *(Figure-1-3)*. Massive-cleaning impairs data-representativeness and challenges model's learning-performance over real-patterns from original-source [35]. Additionally, handling list-data with basic binary-classification simplifies data-processing for computation. However, better to adopt feature-hashing for keeping information for their potential-values, especially nuanced-details. Its binary-handling helps minimise data-processing cost while securing data-completeness [36]. Regarding outlier-scaling, while 1.5x IQR fits for general-practices, pressing z-score to 1.75 is overly strict, though the model's tuning is based on accuracy-matrix [37]. In our case, high-rents possibly resulted from luxury-area but boxplots *(Figure-18)*'s data only ranged to nearly $2000. Overcleaning, therefore, leads to slight-distortion of original-data [38].

Regarding data-analysis, the project relies on traditional statistical-methods. As *task-1(b)* mentioned, ties will lead to deviations while adopting Spearman's-*"p"* in datasets with highly-concentrated values [38]. Adopting additional-methods for cross-checks could enhance interpretations [Ibid, p.44]. However, the project overlooks advanced-methods, particularly SHAP which specifically indicates potential-factors influencing variation of rental-prices, though SHAP performs worse in some case-studies [39], [40]. Furthermore, ANOVA is not ideal to compare addresses alongside other discrete-variables, considering address-subcategories are isolated without proportion/contrast-relationship [41]. Missing degree-of-freedom in ANOVA also restricted statistical-robustness, causing oversimplified-interpretation of varied-factors [42].

For subsequent-adjustments, we proposed to adopt categorical-encoding techniques rather than only meeting data-types requirements for machine-learning models. Apart from relying on model's accuracy and KDE-plots observation, keep and labelled targeted-outliers also help exploring hidden-insights from their distribution-pattern though *"their suitability … is typically unknown"* [43]. With available details of address, spatial-scattering is appreciated for analysing correlation price-variations with their geographical-ties, offering concrete and meaningful-insights for future-analysis [44].

2. Task-2

2.1. Task-2(a)

Task-2(a) designs SQL-database from original CSV-datasets, with minor-amendments for categorisation and 3NF's fulfilment.

```
CREATE TABLE IF NOT EXISTS tenant (
    tenant_id INT PRIMARY KEY AUTO_INCREMENT UNIQUE,
    age INT NOT NULL CHECK (age > 0),
    gender ENUM("MALE", "FEMALE", "OTHERS") NOT NULL,
    is_active BOOLEAN DEFAULT TRUE
);
```

Figure-19: SQL-codes - ENUM and CHECK instructions (table tenant as example)

In 1NF-stage, the proposal addresses duplications and missing-values. New-design setup the new primary-key *(known as *_id)* and restricted to be UNIQUE and NOT NULL with incrementally-numbered [45]. It is more traceable and readable than UUIDs-alternatives with less storage-cost, especially efficient in high-workload [46], [47]. Additionally, the schema standardised inputs in designated data-types, even strict-regulations upon instructions. For instance, *Figure-19* ENUM and CHECK limit possible input-range for ensuring data-consistency and lower-maintenance costs [48].

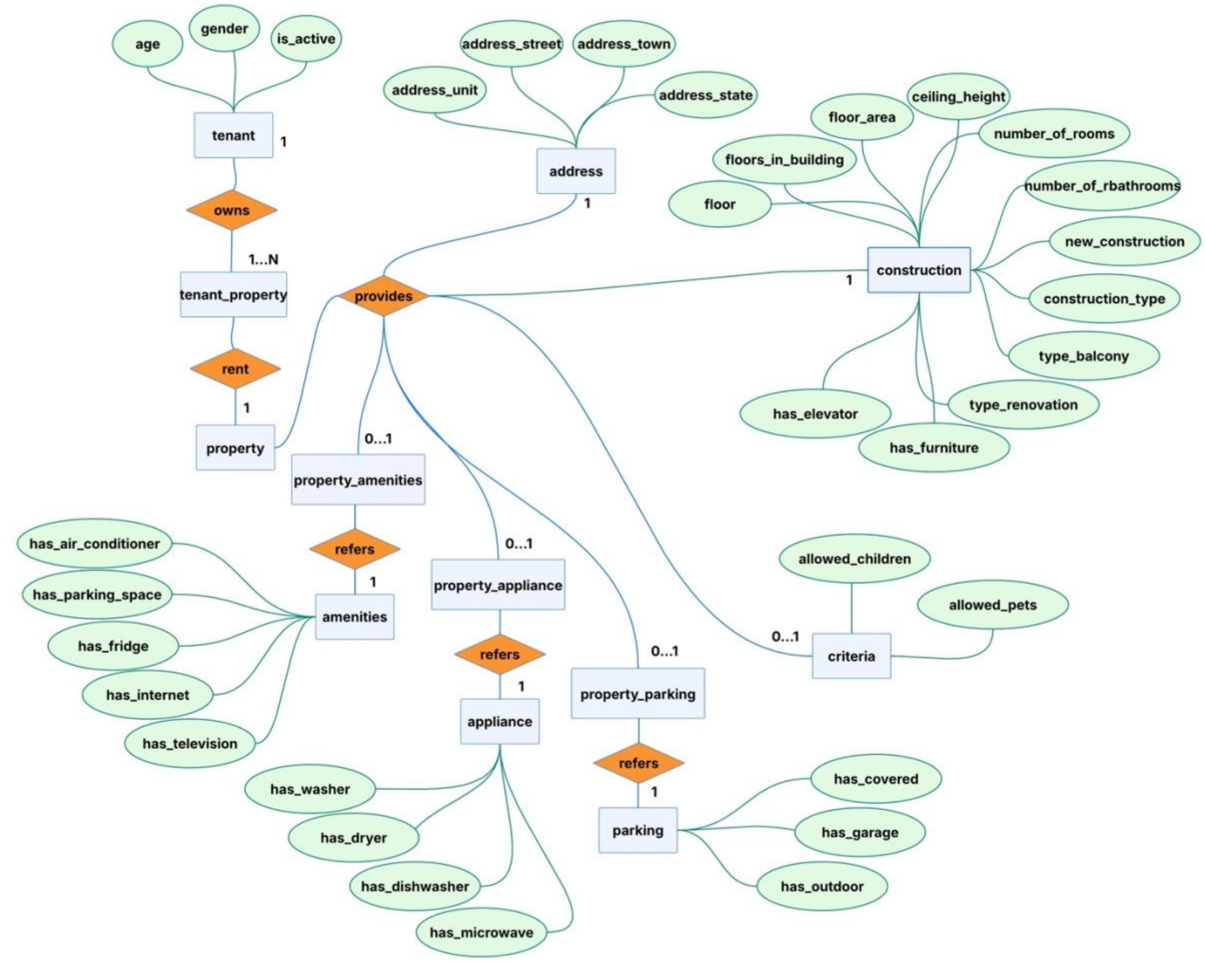

Figure-20: ER-diagram - SQL-database design

2NF-stage requires attributes to rely on their primary-key each table [45, pp.739]. Hence, database-structure *(Figure-20)* can be divided into three-layers - *"tenant"*, *"tenant_property"* and *"property"*, preventing mess of attributes and duplication. Subordinated-tables, such as *"address"*, *"construction"* and *"criteria"* are divided, dedicated for scalable data-storage and supporting subsequent-queries [49]. In *"Address"*, specific input of unit/street/town simplifies processing in *task-1* and better-categorised for machine-learning and analysis. The major-tables also contain the attribute *"is_active"* which labelled current-validity of records, leaving traces for data-history/future-use *(Figure-19)* [50].

```
CREATE TABLE IF NOT EXISTS property_amenities (
    property_id INT,
    amenities_id INT,
    PRIMARY KEY (property_id, amenities_id),
    FOREIGN KEY (property_id) REFERENCES property(property_id),
    FOREIGN KEY (amenities_id) REFERENCES amenities(amenities_id)
);
```

Figure-21: SQL-codes - Foreign-keys demonstration (property_amenities as example)

3NF-stage eliminates transitive-dependency [45, pp.741]. Through foreign-keys implementation, attributes only rely on their primary-key, reducing redundant-data and potential-anomalies. In *"tenant_property"*, foreign-keys of *"tenant_id"* and *"property_id"* helps explaining one-to-many relationship, while list-based tables *("appliance", "amenities", "parking")* fits to describe many-to-many relationship that bridging intermediate-tables *(Figure-21)* [51]. And the adaptors of intermediate-tables strengthened data-consistency with lossless-join relation, overall organisation and effective-maintenance with simple connection [52].

```
SELECT
    property.property_id, address.address_unit, address.address_street,address.address_town,
    address.address_state, tenant_property.price, tenant_property.currency, construction.has_elevator
        FROM property
            JOIN address ON property.address_id = address.address_id
            JOIN construction ON property.construct_id = construction.construct_id
            JOIN tenant_property ON property.property_id = tenant_property.property_id
                WHERE
                    tenant_property.price <= 1000
                        AND construction.has_elevator = TRUE
                        AND tenant_property.currency = "USD"
                        AND tenant_property.is_active = TRUE
                        AND property.is_active = TRUE;
```

property_id	address_unit	address_street	address_town	address_state	price	currency	has_elevator
6	Antarayin 1st blok 10, Tsaghkadzor	Antarayin 1st blok 10	Tsaghkadzor	NULL	1000	USD	1

Figure-22: SQL-codes - Answer Q2b(ii) and result for demonstrating filtering and viewing

3NF demonstrates flexibility in modelling and analysis. *Figure-22* reveals nominalisation improves filtering and viewing based on special attributes, optimising later machine-learning processes with easier data-conversion and feature-engineering. Primary-keys trace data-entries and remove duplication, and foreign keys clarify inter-relationships between tables, improving model-interpretation for learning statistical-patterns [53].

```sql
INSERT INTO tenant
    (age, gender)
        VALUES
            (31, "FEMALE");
SET @tenant_id = LAST_INSERT_ID();

INSERT INTO address
    (address_unit, address_town)
        VALUES
            ("Davidashen 4-th block, Yerevan", "Yerevan");
SET @address_id = LAST_INSERT_ID();

INSERT INTO construction
    (construction_type, new_construction, has_elevator, floors_in_building, floor_area,
    number_of_rooms, number_of_bathrooms, ceiling_height, floor, type_balcony,
    has_furniture, type_renovation)
        VALUES
            ("MONOLITH", TRUE, TRUE, 9, 85,3, 1, 3, 4, "OPEN",NULL, "EURO");
SET @construction_id = LAST_INSERT_ID();
INSERT INTO criteria
    (allowed_children, allowed_pets)
        VALUES
            (TRUE, TRUE);
SET @criteria_id = LAST_INSERT_ID();

INSERT INTO property
    (address_id, construct_id, criteria_id, is_active)
        VALUES
            (@address_id, @construction_id, @criteria_id, TRUE);
SET @property_id = LAST_INSERT_ID();

INSERT INTO tenant_property
    (tenant_id, property_id, purge_date, duration, price, currency, type_utility_payments, is_active)
        VALUES
        (@tenant_id, @property_id, STR_TO_DATE("07/01/2023", "%m/%d/%Y"), "MONTHLY", 1100, "USD", NULL, TRUE);

INSERT INTO amenities
    (has_air_conditioner, has_parking_space, has_fridge, has_internet, has_television)
        VALUES
            (FALSE, FALSE, FALSE, FALSE, FALSE);
SET @amenities_id = LAST_INSERT_ID();
```

```sql
INSERT INTO appliance
    (has_washer, has_dryer, has_dishwasher, has_microwave)
        VALUES
            (FALSE, FALSE, FALSE, FALSE);
SET @appliance_id = LAST_INSERT_ID();

INSERT INTO parking
    (has_covered, has_garage, has_outdoor)
        VALUES
            (FALSE, FALSE, FALSE);
SET @parking_id = LAST_INSERT_ID();

INSERT INTO property_amenities (property_id, amenities_id) VALUES (@property_id, @amenities_id);

INSERT INTO property_appliance (property_id, appliance_id) VALUES (@property_id, @appliance_id );

INSERT INTO property_parking (property_id, parking_id) VALUES (@property_id, @parking_id);
```

Figure-23: SQL-codes - Answer Q2b(i) for demonstrating complicated instructions

However, 3NF-nominalisation brings restrictions in practice. As *Figure-23* showed, it requires multiple and complicated instructions to exercise single-entities. Verification of identifier-layers and multiple attributes are expensive in data-processing and maintenance [54]. Furthermore, with over-normalising schema, fragmented tables lead to higher-complexity with excessive JOIN tables with worse-effectiveness and readability, increasing technical-barriers on processors/analysts *(Figure-22)* [55].

The below section demonstrates SQL-code exercises from *task-2(a)(i-iii).* Ten sample-entries are imported for demonstration *(Figure-24)*.

```sql
SELECT * FROM tenant_property;
```

rental_id	tenant_id	property_id	purge_date	duration	price	currency	type_utility_payme...	is_active
1	1	1	2025-06-26 14:34:11	MONTHLY	1200	USD	ALL	1
2	2	2	2025-06-26 14:34:11	MONTHLY	1300	USD	ALL	1
3	3	3	2025-06-26 14:34:11	MONTHLY	1100	USD	ALL	1
4	4	4	2025-06-26 14:34:11	MONTHLY	1050	USD	ALL	1
5	5	5	2025-06-26 14:34:11	MONTHLY	1500	USD	ALL	1
6	6	6	2025-06-26 14:34:11	MONTHLY	1000	USD	ALL	1
7	7	7	2025-06-26 14:34:11	MONTHLY	1800	USD	ALL	1
8	8	8	2025-06-26 14:34:11	MONTHLY	1400	USD	ALL	1
9	9	9	2025-06-26 14:34:11	MONTHLY	1250	USD	ALL	1
10	10	10	2025-06-26 14:34:11	MONTHLY	1350	USD	ALL	1

Figure-24: SQL-code - display initialised dataset (selected 10-entries from original)

For inserting new-entries, INSERT data-rows appends data to each table separately. And user-defined variables are references for connecting relationships among tables, enabling relevant-data to be connected within complex database-structure *(Figure-25)*. The 11th-row is therefore appended successfully.

```
INSERT INTO tenant
    (age, gender)
        VALUES
            (31, "FEMALE");
SET @tenant_id = LAST_INSERT_ID();

INSERT INTO address
    (address_unit, address_town)
        VALUES
            ("Davidashen 4-th block, Yerevan", "Yerevan");
SET @address_id = LAST_INSERT_ID();

INSERT INTO construction
    (construction_type, new_construction, has_elevator, floors_in_building, floor_area,
    number_of_rooms, number_of_bathrooms, ceiling_height, floor, type_balcony,
    has_furniture, type_renovation)
        VALUES
            ("MONOLITH", TRUE, TRUE, 9, 85,3, 1, 3, 4, "OPEN",NULL, "EURO");
SET @construction_id = LAST_INSERT_ID();
INSERT INTO tenant
    (age, gender)
        VALUES
            (31, "FEMALE");
SET @tenant_id = LAST_INSERT_ID();

INSERT INTO address
    (address_unit, address_town)
        VALUES
            ("Davidashen 4-th block, Yerevan", "Yerevan");
SET @address_id = LAST_INSERT_ID();

INSERT INTO construction
    (construction_type, new_construction, has_elevator, floors_in_building, floor_area,
    number_of_rooms, number_of_bathrooms, ceiling_height, floor, type_balcony,
    has_furniture, type_renovation)
        VALUES
            ("MONOLITH", TRUE, TRUE, 9, 85,3, 1, 3, 4, "OPEN",NULL, "EURO");
SET @construction_id = LAST_INSERT_ID();
```

```sql
INSERT INTO tenant
    (age, gender)
        VALUES
            (31, "FEMALE");
SET @tenant_id = LAST_INSERT_ID();

INSERT INTO address
    (address_unit, address_town)
        VALUES
            ("Davidashen 4-th block, Yerevan", "Yerevan");
SET @address_id = LAST_INSERT_ID();

INSERT INTO construction
    (construction_type, new_construction, has_elevator, floors_in_building, floor_area,
    number_of_rooms, number_of_bathrooms, ceiling_height, floor, type_balcony,
    has_furniture, type_renovation)
        VALUES
            ("MONOLITH", TRUE, TRUE, 9, 85,3, 1, 3, 4, "OPEN",NULL, "EURO");
SET @construction_id = LAST_INSERT_ID();
```

rental_id	tenant_id	property_id	purge_date	duration	price	currency	type_utility_payme...	is_active
1	1	1	2025-06-26 14:34:11	MONTHLY	1200	USD	ALL	1
2	2	2	2025-06-26 14:34:11	MONTHLY	1300	USD	ALL	1
3	3	3	2025-06-26 14:34:11	MONTHLY	1100	USD	ALL	1
4	4	4	2025-06-26 14:34:11	MONTHLY	1050	USD	ALL	1
5	5	5	2025-06-26 14:34:11	MONTHLY	1500	USD	ALL	1
6	6	6	2025-06-26 14:34:11	MONTHLY	1000	USD	ALL	1
7	7	7	2025-06-26 14:34:11	MONTHLY	1800	USD	ALL	1
8	8	8	2025-06-26 14:34:11	MONTHLY	1400	USD	ALL	1
9	9	9	2025-06-26 14:34:11	MONTHLY	1250	USD	ALL	1
10	10	10	2025-06-26 14:34:11	MONTHLY	1350	USD	ALL	1
11	11	11	2023-07-01 00:00:00	MONTHLY	1100	USD	NULL	1

Figure-25: task-2b(ii) - Inserting new entry covering all relevant attributes

For querying items with specific-requirements, SELECT methods are used for covering designated-columns from various tables, combining into result-table with JOIN method and WHERE keyword for matching logical-conditions. One matched result is filtered successfully *(Figure-26)*.

```sql
SELECT
    property.property_id, address.address_unit, address.address_street,address.address_town,
    address.address_state, tenant_property.price, tenant_property.currency, construction.has_elevator
        FROM property
            JOIN address ON property.address_id = address.address_id
            JOIN construction ON property.construct_id = construction.construct_id
            JOIN tenant_property ON property.property_id = tenant_property.property_id
                WHERE
                    tenant_property.price <= 1000
                        AND construction.has_elevator = TRUE
                        AND tenant_property.currency = "USD"
                        AND tenant_property.is_active = TRUE
                        AND property.is_active = TRUE;
```

property_id	address_unit	address_street	address_town	address_state	price	currency	has_elevator
6	Antarayin 1st blok 10, Tsaghkadzor	Antarayin 1st blok 10	Tsaghkadzor	NULL	1000	USD	1

Figure-26: task-2b(ii) - Querying items for less than $1000 USD with elevator included

For querying average-price in specific currency, extracted currency and calculated-price columns, and GROUP BY keywords forms new pivot-tables and ORDER BY keyword for sorting. Only group of USD records $1277.27 on average *(Figure-27)*.

```
SELECT currency, ROUND(AVG(price), 2) AS average_price
    FROM tenant_property
        WHERE is_active = TRUE
            GROUP BY currency
                ORDER BY average_price DESC;
```

currency	average_price
USD	1277.27

Figure-27: task-2c(iii) - Extract Average Price for Each Currency

2.2. Task-2(b)

Task-2 outlines requirements for system-scalability and data-management. With the scenario of global renting-agent, we assumed that frequent and synchronous inquiries come from locals themselves, while headquarters regularly audits with cross-regional access. Foreseeing rapid data-size growth, it highlights the unpredicted demands on system-capacity, and fault-prevention, elastic-expansion and data-consistence measures are highly-appreciated. As real-time alerts have been specified, we expected the system to react upon pre-set criteria but immediate responses with low-latency are crucial.

Thus, cloud-based solutions *(AWS-S3, AWS-EMR and AWS-Lambda)* are selected which offer better scalability and elasticity than traditional Hadoop-ecosystem, following advantages of geographical-distribution and rapid-response. Indeed, industries gradually flavour modern-alternative regarding simplicity and cost-efficiency, with better-performances resulting in comparative-studies [56], [57]. As multiple cloud-service providers offer similar technologies on handling big-data, AWS is used for further explanation.

Regarding data-storage, S3 *(cloud-based data-storage)* offers comparability, better elasticity and data-consistency. But Hadoop-processing requires longer-time to load large-batches, revealing its unsuitability towards real-time rapid-response requests [58]. Compared to HDFS's horizontal-scaling, clusters of cloud-servers automatically scale upon data-size changes, without service-downtime/interruption and professional-knowledge on managing name-node's metadata [59], [60]. Furthermore, HDFS's distributed-storage stores data-blocked copies across various data-nodes separately [59]; however, AWS takes advantage of backup files with their data-centre networks continuously and inter-regionally [61]. Thus, users could get data from nearest data-centre and no longer wait for rebuilding copies between name-nodes and foreign data-nodes [62], [63].

S3 is preferred over hybrid-approaches though its higher cost. It prevents massive-spending over initial facilities and long-term maintenance; but future growing data-size and frequent requests can lead to heavy-cost, especially over-budget with *"pay-as-you-go"* plans during peak periods [64], [65]. Although hybrid-approach could optimise the cost with insignificant data, it requires advanced techniques to align data-consistency one another, further causing higher complexity to server-operation and maintenance [66]. Property-data is considered as mainly region-based and rarely being foreign requested. Potentially, optimising performance and resource-management with temperature-based models, plus reduction on cross-regional requests [67].

Regarding data-processing, adopting EMR and Lambda shows better performance on fault-tolerance and rapid-response. EMR's computation is automatically-scalable based on the task's workload, allowing distributed-tasks for processing synchronously without manual adjustment or system interruption [68]. Additionally, the serverless Lambda provides low-latency execution with streaming-data which specifically live-alerts and minimised computation-consumption [69]. Though urgent-alerts are rare in property-renting, Lambda awakens from idle-status which causes several-seconds delay of alerting [70]. Excepting batch-processing delay, traditional options like Apache-Spark reduce overhead in MapReduce but still require manual-control over node-clusters which is less applicable for time-sensitive situations, while cloud-native benefits on global-scalability [71], [72].

Although EMR supports Spark, its auto-scaling is restricted by the AWS built-in foundation. Unlike mature-integration between traditional HDFS, Spark/MapReduce and YARN, EMR's PAAS-nature does not offer low-level customisation, namely YARN-scheduling which is possibly causing resource-contention amidst high-concurrent procession for the agent-company [73] EMR is struggling to allocate tasks/resources efficiently with delay/degraded-service, while concurrent modelling/analysing tasks are ongoing [74]. Nevertheless, EMR's simplified architecture, automotive-practice and elastic-scalability become compatible for unpredicted needs of fast-business and data-size growth *(up to tens of megabytes)* in future [75]. Lambda's event-driven nature further fulfilled real-time process which Hadoop-ecosystem hardly-managed [76].

3. Task-3

With front-end view, moral and legal issues of excessive data-collection are critical. Instead of analysing tenant behaviours, data like *"gender"* and *"age"* are less relevant for price-correlation analysis *(in task-1)*. Rich datasets help analysis but expand data-breach impact and more risk-management investment [77]. Excessive data-collection contradicts data-minimisation, raising censorship concerns and public-distrust [78], [79]. It reduces user-engagement and triggers hate-feedbacks, harming model performance and accuracy [80], [81].

For online-survey, only essential data is needed for business goals with optional-fields for leaving fewer traces [Ibid, pp.686-687]. Front-end developers can mask sensitive-data before transmission by adopting libraries of *"bcrypt"* and *"uuid"* [82]. In long-run, management should evaluate business-goals and data-collecting practices. Under GDPR, companies must strengthen data-retention-policies to ensure data necessity with historical-data disposal [83]. Automation of flagging unused-data or scheduling data-cleaning with scripts helps comply with requirements in efficient way [84].

Front-end manages user-interface, but data-protection depends on backend processing and organisation [85]. Hashing is irreversible and tool *"crypto"* also supports encryption, protecting data-transmission against hacking alongside HTTPS [86], [87]. But staff incentives and awareness need proper work-culture with time and resources [88].

Reviewing back-end, overfitting--models threaten privacy via model-inversion and membership-inference attacks [89]. As classifier/regressor this task required, unique, extreme and recognisable entries could be maliciously-exposed [90]. Current surveys/findings highlight how model-inversion endangers data-privacy in different perspectives, and exploring potential security-countermeasures [91]. Enterprises must defend data from cyber-attacks and leakage, as negligence leads to legal-liabilities and loss amidst regulations [92].

Basically, risk induction with model-output processing, namely appending addictive noise. Through reducing model-inversion risk cost-effectively [93]. Huang's findings demonstrate how multiple noise-addiction strategies enhance data-security against membership-inference, by avoiding training test-data directly [94]. As for the model, technique of preventing overfitting by approaches, such as fundamentals like memorising-basis with restricted training-process / adding adversarial-prompts for risk-mitigation [95]. Long-term, management's regular-reviews of data-necessity and model-sensitivity *(task-1)* helps supervise protection and model-consistency with timely-adjustment [96].

Addictive-noise mitigates overly-memorisation for better stability and accuracy but impacts negatively towards model's precision if noise-level is high [97]. This issue relates to underfitting caused by insufficient/oversimplified models for learning data-patterns; oppositely, resulted in poorer performance [98]. However, dilemma between protection and precision, legal-risks and benefits, involves internal-political/managerial considerations that exceed technical extent [99].

SQL-injection is a common database-threat. Though without login, blank input-fields and URL-parameters can be hacked by malicious-coding [100]. Chronological SQL-data indices leave vulnerabilities, as input matches identity with tautologies, union-query, and blind-injection attacks [101], [102]. Without appropriate access-settings, hackers override databases with system-instructions, obtaining privacy for further attacks causing loss [103]. Despite no technical-standard legally enforced, enterprises may contravene law through negligence, given legal-responsibilities on reliable data-protection [104], [105].

Yaswanthraj et al. suggested formulating SQL-instructions through ORM for low cost and high-efficiency with automatic/strict-coding standard [106]. In coding-practice, stored credentials in config-files or encrypted media with one-off password help preventing exposure [107]. Long-term, technical support for protection, maintenance and troubleshooting, and data-protection-officers for management, training and supervision for compliance, would safeguard data-privacy with business-sustainability [108], [109]. CISSP and Ashbaugh [110] further suggested appending risk-management to software-development-lifecycle, applying assessment and management for eliminating vulnerabilities during development.

Nevertheless, ORM is not an all-rounded solution, considering its restrictions on *“huge learning curve, performance issues, and single-platform compatibility”* [111] Though technical-techniques and practices are well-covered, programming failures still be expensive for later troubleshooting/maintenance, overloading support and data-security and risking possible-penalties [112].

Overall, technical solutions and management are significant, but stable work-cultures - genuinely respecting data-privacy is crucial, rather than mere legal-regulations [113], [114]. Otherwise, hidden vulnerabilities and data privacy remain unresolved.

Endnote

[1] O. Marban et al., “A Data Mining and Knowledge Discovery Process Model,” in J. Ponce and A. Karahoca (eds), *Data Mining and Knowledge Discovery in Real Life Applications*, Rijeka: InTech, 2012, pp.3.

[2] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048, Dec. 2019, doi: https://doi.org/ 10.1109/tkde.2019.2962680.

[3] J. Abasova, P. Tanuska and S. Rydzi, “Big Data—Knowledge Discovery in Production Industry Data Storages - Implementation of Best Practices,” *Applied Sciences*, vol. 11, no. 16, pp. 7648, Aug. 2021, doi: https://doi.org/10.3390/app11167648.

[4] P. Chapman et al., “The CRISP-DM process model,” pp. 9, 1999. [Online]. https://keithmccormick.com/wp-content/uploads/CRISP-DM%20No%20Brand.pdf.

[5] Supercasa, “Rents Could Rise by Up to 11% in 2025 in Specific Cases - SUPERCASA,” 2025..https://supercasa.pt/en-gb/noticias/rents-could-rise-by-up-to-11-in-225-in-specific-cases/n6385 (accessed Jun. 19, 2025).

[6] O. F. Ayilara et al., “Impact of Missing Data on Bias and Precision When Estimating Change in Patient-Reported Outcomes from a Clinical Registry,” *Health Qual Life Outcomes*, vol. 17, no. 1, pp. 106, 2019, doi: https://doi.org/10.1186/s12955-019-1181-2.

[7] C. Hanig, M. Schierle, D. Trabold, “Comparison of Structured vs. Unstructured Data for Industrial Quality Analysis,” in *Proceedings of the World Congress Engineering and Computer Science Vol. 1,* October 20-22, 2010, San Francisco, United States, pp.7, [Online]. https://www.iaeng.org/publication/WCECS2010/WCECS2010_pp432-438.pdf.

[8] F. Ododo and N. Addotey, “Understanding the Influence of Outliers on Machine Learning Model Interpretability,” *International Journal of African Sustainable Development Research, vol. 7, no. 2, pp.51, February 2005, doi:* https://doi.org/*10.70382/tijasdr.*

[9] F. Ridzuan and W. Zainon, “A Review on Data Cleansing Methods for Big Data,” *Procedia Computer Science*, vol. 161, pp. 731-732, Jan. 2019, doi:https://doi.org/10.1016/j.procs.2019.11.177.

[10] J. M. Hellerstein, “Qualitative Data Cleaning for Large Databases,” United Nations Economic Commission for. Europe (UNECE), pp. 1-42, February 2008. [Online]. Accessed: https://dsf.berkeley.edu/jmh/papers/cleaning-unece.pdf.

[11] “Top Ten Ways to Clean Your Data - Microsoft Support,” support.microsoft.com. https://support.microsoft.com/en-gb/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19. (accessed Jun. 19, 2025).

[12] C. Kuzudisli et al., “Review of Feature Selection Approaches based on Grouping of Features,” *PeerJ,* vol. 11, no. e15666, pp.26, July 2023, doi: https://doi.org/10.7717/peerj.15666.

[13] B. Dit et al., "Can Better Identifier Splitting Techniques Help Feature Location?," *2011 IEEE 19th International Conference on Program Comprehension*, Kingston, ON, Canada, 2011, pp. 19, doi: 10.1109/ICPC.2011.47.

[14] N. B. Konda, “The Impact of Data Preprocessing on Data Mining Outcomes,” *World Journal of Advanced Research and Reviews*, vol. 15, no. 3, pp. 542, Sep. 2022, doi: https://doi.org/10.30574/wjarr.2022.15.3.0931.

[15] P. Pranav, A. Patel and S. Jain, eds., *Machine Learning in Healthcare and Security: Advances, Obstacles, and Solutions*. Boca Raton, CRC Press, 2024, pp.7.

[16] J. Kirchner, A. Heberle and W. Löwe, "Classification vs. Regression - Machine Learning Approaches for Service Recommendation Based on Measured Consumer Experiences," *2015 IEEE World Congress on Service*s, New York, NY, USA, 2015, pp. 278, doi: https://doi.org/10.1109/SERVICES.2015.49.

[17] Q. Chen and J. Qi, “How Much Should We Trust R2 and Adjusted R2 : Evidence from Regressions in Top Economics Journals and Monte Carlo Simulations,” *Journal of Applied Economics*, vol. 26, no. 1, pp.1, May 2023, doi: https://doi.org/10.1080/15140326.2023.2207326.

[18] G. Linoff and M. Berry, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship*. Indianapolis, Wiley Publishing, 2011, pp. 219.

[19] D. Powers, “Evaluation: From Precision , Recall and F-Measure to Roc, Informedness, Markedness & Correlation,” *arXiv (Cornell University)*, pp.37, Jan. 2020, doi: https://doi.org/10.48550/arxiv.2010.16061.

[20] M. Jiang et al., “Random Forest Clustering for Discrete Sequences,” *Pattern Recognition Letters*, vol. 174, pp. 150, Sep. 2023, doi: https://doi.org/10.1016/j.patrec.2023.09.001.

[21] B. Gregorutti, B. Michel and P. Saint-Pierre, “Correlation and variable importance in random forests,” *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, Mar. 2016, doi: https://doi.org/10.1007/s11222-016-9646-1.

[22] G. Louppe et al., “Understanding Variable Importances in Forests of Randomized Trees,” *Neural Information Processing Systems*, vol. 26, pp.432, Dec. 2013. [Online]. Available: http://orbi.ulg.ac.be/bitstream/2268/155642/1/louppe13.pdf.

[23] J. F. Plante and M. Radatz, “On the Capability of Classification Trees and Random Forests to Estimate Probabilities,” J*ournal of Statistical Theory and Practice*, vol. 18, no. 2, pp. 24-25, Apr. 2024, doi: https://doi.org/10.1007/s42519-024-00376-5.

[24] A. McCluskey and A. Lalkhen, “Statistics II: Central Tendency and Spread of Data, ” *Continuing Education in Anaesthesia, Critical Care & Pain*, vol.7, no. 2, pp.127, 2007.

[25] G. P. Quinn and M. J. Keough, *Experimental Design and Data Analysis for Biologists.* Cambridge: Cambridge University Press, 2002, pp. 61.

[26] A. Mazarei et al., “Online Boxplot Derived Outlier Detection,” *International Journal of Data Science and Analytics*, May 2024, pp.85, doi: https://doi.org/10.1007/s41060-024-00559-0.

[27] “Spearman’s Rank-Difference Coefficient of Correlation,” in T. W. MacFarland and J. M. Yates, eds., *Introduction to Nonparametric Statistics for the Biological Sciences Using R.* Cham: Springer Nature, 2016, pp. 249, doi: https://doi.org/10.1007/978-3-319-30634-6_8.

[28] M. T. Puth, M. Neuhäuser and G. D. Ruxton, “Effective Use of Spearman’s and Kendall’s

Correlation Coefficients for Association Between Two Measured Traits," *Animal Behaviour*, vol. 102, pp. 82, Feb. 2015, doi: https://doi.org/10.1016/j.anbehav.2015.01.010.

[29] T. Boonupara, P. Udomkun and P. Kajitvichyanukul, "Quantitative Analysis of Atrazine Impact on UAV-Derived Multispectral Indices and Correlated Plant Pigment Alterations: A Heatmap approach," *Agronomy*, vol. 14, no.4, 814, pp.17, Apr. 2024, doi: https://doi.org/10.3390/agronomy14040814.

[30] L. Kennedy-Shaffer, "Before p < 0.05 to Beyond p < 0.05: Using History to Contextualize p-Values and Significance Testing," *The American Statistician*, vol. 73, no. sup1, pp. 89, Mar. 2019, doi: https://doi.org/10.1080/00031305.2018.1537891.

[31] M. Krzywinski and N. Altman, "Visualizing Samples with Box Plots," *Nat Methods*, vol. 11, no. 2, pp. 119, Jan. 2014, doi: https://doi.org/10.1038/nmeth.2813.

[32] K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton, CRC Press, 2016, pp.14.

[33] T. K. Kim, "Understanding One-way ANOVA Using Conceptual Figures," *Korean Journal of Anesthesiology*, vol. 70, no. 1, pp. 22, 2017, doi: https://doi.org/10.4097/kjae.2017.70.1.22.

[34] Q. Liu and L. Wang, "T-Test and ANOVA for Data with Ceiling and/or Floor Effects," *Behavior Research Methods*, vol. 53, pp. 264, 2021, doi: https://doi.org/10.3758/s13428-020-01407-2.

[35] T. Dasu and J. M. Loh, "Statistical distortion: Consequences of data cleaning," *arXiv (Cornell University)*, pp. 1681, Jan. 2012, doi: 10.48550/arxiv.1208.1932.

[36] P. Duboue, *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge: Cambridge University Press, 2020, pp. 100.

[37] J. Howells, *Data Science for Decision Makers: Enhance your Leadership Skills with Data Science and AI Expertise*. Birmingham: Packt Publishing, 2024, pp. 55.

[38] C. Salgado et al., "Noise Versus Outliers, " in MIT Critical Data, eds., *Secondary Analysis of Electronic Health Records*. Cham: Springer, 2016, pp.281.

[39] L. Wu, "A Review of the Transition from Shapley Values and SHAP Values to RGE," *Statistics*, pp.20, Apr. 2025, doi: https://doi.org/10.1080/02331888.2025.2487853

[40] H. Wang et al., "Feature Selection Strategies: A Comparative Analysis of SHAP-value and Importance-based Methods," *J Big Data*, vol. 11, no. 44, pp. 15, 2024. doi: https://doi.org/10.1186/s40537-024-00905-w.

[41] E. Acquiles et al., "The Effect of Spatial Lag on Modeling Geodetic Covariates Using Analysis of Variance," *Applied Geomatrics*, vol. 16, pp. 786-787, doi: https://doi.org/10.1007/s12518-024-00579-2.

[42] J. G. Eisenhauer, "Degrees of Freedom in Statistical Inference," in M. Lovric, eds, *International Encyclopedia of Statistical Science*. Berlin and Heidelberg: Springer, 2011, pp. 365, doi: https://doi.org/10.1007/978-3-642-04898-2_24.

[43] G. O. Campos et al., "On the Evaluation of Unsupervised Outlier Detection: Measures,

Datasets, and an Empirical Study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 892, Jan. 2016, doi: https://doi.org/10.1007/s10618-015-0444-8.

[44] S. Basu and T. Thibodeau, "Analysis of Spatial Autocorrelation in House Prices," *Journal of Real Estate Finance and Economics*, vol 17, no. 1, pp. 62-63, 1998. doi: https://doi.org/10.1023/A:1007703229507.

[45] T. Haloin and T. Morgan, *Information Modeling and Relational Databases, Second Edition.* Burlington, Elsevier Inc., 2008, pp. 737.

[46] J. Huang et al., "Efficient Auto-Increment Keys Generation for Distributed Log-Structured Storage Systems," in H. Hacid et al., eds, *Web Information Systems Engineering – WISE 2018. WISE 2018 Lecture Notes in Computer Science*, vol 11234. Cham: Springer, pp. 225, doi: https://doi.org/10.1007/978-3-030-02925-8_16.

[47] D. Farrell, *The Well-Grounded Python Developer: How the Pros Use Python and Flask*. Shelter Island: Manning Publication, 2023, pp. 200.

[48] H. Zhang et al., "Checking Enforcement of Integrity Constraints in Database Applications Based on Code Patterns," *The Journal of System and Software*, vol. 84, no. 12, pp. 2253, 2011, doi: https://doi.org/10.1016/j.jss.2011.06.044.

[49] T. Taipalus, "On the Effects of Logical Database Design on Database Size, Query Complexity, Query Performance, and Energy Consumption," *arXiv.org*, pp. 1, Jan. 13, 2025, doi: https://doi.org/10.48550/arXiv.2501.07449.

[50] "Stack Overflow," *Stack Overflow*, Sep. 09, 2011. https://stackoverflow.com/questions/7366849/implementing-soft-delete-with-minimal-impact-on-performance-and-code (accessed Jun. 26, 2025).

[51] V. Rainardi, *Building a Data Warehouse: With Example in SQL Server*. Berkeley: Apress, 2008, pp. 109.

[52] S. Wong, M. Tsou and X. Jing, "Database Schema Design In The Relational Model," *INFOR: Information Systems and Operational Research*, vol. 23, pp. 3, 1985. doi: https://doi.org/10.1080/03155986.1985.11731941.

[53] L. Jiang and F. Naumann, "Holistic Primary Key and Foreign Key Detection," *Journal of Intelligent Information Systems,* vol. 54, pp. 460, 2020, doi: https://doi.org/10.1007/s10844-019-00562-z.

[54] M. Albarak et al., "Managing Technical Debt in Database Normalization," *IEEE Transactions on Software EngineeringI, vol. 48, no. 3, pp. 755-756, March 2022, doi:* https://doi.org/*10.1109/TSE.2020.3001339.*

[55] J. Garmany, J. Walker and T. Clark, *Logical Database Design Principles*. Boca Raton: CRC Press, 2005, pp. 140.

[56] R. S. Koppula and Satsyil Corp, "Comparative Analysis of Big Data Storage Strategies: Hadoop Distributed File System (HDFS) vs. Cloud-based Solutions," Journal of Scientific and Engineering Research, pp. 262-263, Dec. 2020. [Online]. Available: https://jsaer.com/download/vol-7-iss-12-2020/JSAER2020-7-12-258-263.pdf

[57] H. K. Chaubey et al., "Cloud Versus Local: Performance Evaluation of Multi-node Hadoop Clusters Using HiBench Benchmarks," in *Lecture Notes in Networks and Systems*, 2024,

pp. 24. doi: https://doi.org/10.1007/978-3-031-73110-5_2.

[58] M. T. Gabdullin, Y. Suinullayev, Y. Kabi, J. W. Kang, and A. Mukasheva, “Comparative Analysis of Hadoop and Spark Performance for Real-time Big Data Smart Platforms Utilizing IoT Technology in Electrical Facilities,” *Journal of Electrical Engineering and Technology*, vol. 19, no. 7, pp. 4595-4596, Jun. 2024, doi: https://doi.org/10.1007/s42835-024-01937-1.

[59] Y. Kalmukov, M. Marinov, T. Mladenova, and I. Valova, “Analysis and experimental study of HDFS performance,” TEM Journal, pp. 813, May 2021, doi: 10.18421/tem102-38.

[60] S. Alharthi et al., “Auto-Scaling Techniques in Cloud Computing: Issues and Research Directions,” *Sensors*, vol. 24, no. 5551, pp. 5, Aug. 2024, doi: https://doi.org/10.3390/s24175551.

[61] “Creating backup copies across AWS Regions - AWS Backup.” https://docs.aws.amazon.com/aws-backup/latest/devguide/cross-region-backup.html. (Accessed Jun. 23, 2025).

[62] M. Kumar and P. Latha, “Replica Management and High Availability in Hadoop Distributed File System (HDFS),” *International Journal of Engineering Research and Technology*, vol. 3, no. 18, pp. 4, Apr. 2018, [Online]. Available: https://www.ijert.org/research/replica-management-and-high-availability-in-hadoop-distributed-file-system-hdfs-IJERTCONV3IS18020.pdf.

[63] F. Karamimirazizi et al., “Data Replication Methods in Cloud, FOG, and Edge Computing: A Systematic Literature review,” *Wireless Personal Communications*, vol. 135, no. 1, pp. 555, Mar. 2024, doi: https://doi.org/10.1007/s11277-024-11082-7.

[64] R. Fazul, O. Mendizabal and P. Barcelos, “Analyzing the Stability, Efficiency, and Cost of a Dynamic Data Replica Balancing Architecture for HDFS,” *Annals of Telecommunications,* pp. 14, Apr. 2025, doi: https://doi.org/10.1007/s12243-025-01093-1.

[65] C. Chen et al., “Cost Optimization for Serverless Edge Computing with Budget Constraints Using Deep Reinforcement Learning,” *arXiv (Cornell University)*, Jan. 2025, doi: https://doi.org/10.48550/arxiv.2501.12783.

[66] S. Khan et al., “Challenges and Their Practices in Adoption of Hybrid Cloud Computing: An Analytical Hierarchy Approach,” *Security and Communication Networks*, vol. 2021, no. 1, pp. 1, Sept 2021, doi: https://doi.org/10.1155/2021/1024139.

[67] S. Ma et al., “A Hierarchical Storage Mechanism for Hot and Cold Data Based on Temperature Model,” in C. Strauss et al., eds, *Database and Expert Systems Applications*. Cham: Springer Nature, 2024, pp. 155-156.

[68] “Use Amazon EMR cluster scaling to adjust for changing workloads - Amazon EMR,” *Amazon AWS Documentaiton*, https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-scale-on-demand.html. (accessed Jun. 23, 2025)

[69] W. Song et al., “Sponge: Fast Reactive Scaling for Stream Processing with Serverless Frameworks,” in *Proceedings of the 2023 USENIX Annual Technical Conference*, Boston, MA, United States, July 10-12, 2023, pp. 312.

[70] B. Balakrish, “Optimizing AWS Lambda Cold Starts Through Priming: A Technical Exploration,” *International Journal of Computer Engineering and Technology (IJCET), vol. 14, no. 3, pp. 140, 2023,* [Online]. Accessed: *https://iaeme.com/Home/issue/IJCET?Volume=14&Issue=3.*

[71] P. Nghiem, “Best Trade-Off Point Method for Efficient Resource Provisioning in Spark,” *Algorithm*, vol. 11, no.12, pp. 2, 2018, doi: https://doi.org/10.3390/a11120190.

[72] “Real-Time Data Processing Tools Compared | TMA Solutions,” *TMA Solutions*, 2022. https://www.tmasolutions.com/insights/real-time-data-processing-tools-compared (accessed Jun. 26, 2025).

[73] “Configure Hadoop YARN CapacityScheduler on Amazon EMR on Amazon EC2 for Multi-tenant Heterogeneous Workloads | Amazon Web Services,” *Amazon Web Services*, Aug. 16, 2022. https://aws.amazon.com/blogs/big-data/configure-hadoop-yarn-capacityscheduler-on-amazon-emr-on-amazon-ec2-for-multi-tenant-heterogeneous-workloads/ (accessed Jun. 23, 2025).

[74] V. Vyas et al., “Managed Resource Scaling in Amazon EMR,” presented at *SIGMOD-Companion ‘25*, Berlin, Germany, June 22-27, 2025, pp. 4, [Online]. Accessed: https://www.amazon.science/publications/managed-resource-scaling-in-amazon-emr.

[75] C. Rosenau, K. Sandkuhl and B. Nast, “The Role of Business Capabilities for Future Viability in Enterprise Architecture - A Structured Literature Review,” in *Companion Proceedings of the 17th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling Forum*, Stockholm, Sweden, December 3-5, 2024, pp.8.

[76] C. Lekkala, “Leveraging Lambda Architecture for Efficient Real-Time Big Data Analytics,” European Journal of Advances in Engineering and Technology, vol. 7, no. 2, 200, pp. 63.

[77] F. Pino, “The Microeconomics of Data - a Survey,” *Journal of Industrial and Business Economics*, 2022, pp. 658, doi: https://doi.org/10.1007/s40812-022-00220-6.

[78] D. Shanmugam, et. al. “Learning to Limit Data Collection via Scaling Laws: A Computational Interpretation for the Legal Principle of Data Minimization,” *2022 ACM Conference on Fairness, Accountability, and Transparency,* Jun. 2022, pp. 840, doi: https://doi.org/10.1145/3531146.3533148.

[79] N. M. Richards, “The Dangers of Surveillance,” *Harvard Law Review,* vol. 126, no. 7, pp.126, May 2013.

[80] B. Kazansky, “‘It Depends on Your Threat Model’: the Anticipatory Dimensions of Resistance to Data-driven Surveillance,” *Big Data & Society*, vol. 8, no. 1, pp. 205395172098555, 2021, pp. 8-9, doi: https://doi.org/10.1177/2053951720985557.

[81] T. Tryfonas, *Human Aspects of Information Security, Privacy and Trust*. 2017, pp. 683, doi: https://doi.org/10.1007/978-3-319-58460-7.

[82] O. Kulyk, B. M. Reinheimer and M. Volkamer, “Sharing Information with Web Services - A Mental Model Approach in the Context of Optional Information,” in *Lecture notes in Computer Science,* 2017, pp. 675–690, pp. 683-684, doi: https://doi.org/10.1007/978-3-319-58460-7_46.

[83] ICO, “Disposal and Deletion,” *ICO*. https://ico.org.uk/for-organisations/advice-and-services/audits/data-protection-audit-framework/toolkits/records-management/disposal-and-deletion/. 02/06/2025. (accessed: Jun. 19, 2025).

[84] V. Sharma, “The Evolution of Robotic Process Automation in Human Resources: From Recruitment to Retention,” *International Journal of Multidisciplinary Research and Growth Evaluation*, vol. 5, no. 5, pp. 1056, Jan. 2024, doi: https://doi.org/10.54660/.ijmrge.2024.5.5-1053-1058.

[85] “Front End vs Back End - Difference Between Application Development - AWS,” Amazon Web Services, Inc. https://aws.amazon.com/compare/the-difference-between-frontend-and-backend/

[86] “What is Crypto Module in Node.js and How It Is Used ?,” *GeeksforGeeks*, Jun. 14, 2024. https://www.geeksforgeeks.org/what-is-crypto-module-in-node-js-and-how-it-is-used/ (acessed: Jun. 19, 2025).

[87] M. Husák, M. Čermák, T. Jirsík, and P. Čeleda, “HTTPS Traffic Analysis And Client Identification Using Passive SSL/TLS fingerprinting,” *EURASIP Journal on Information Security*, vol. 2016, no. 1, Feb. 2016, pp. 12, doi: https://doi.org/10.1186/s13635-016-0030-7.

[88] E. M. Power, “Developing a Culture of Privacy: a case study,” *IEEE Security & Privacy,* vol. 5, no. 6, pp. 58–60, Nov. 2007, pp. 60, doi: https://doi.org/10.1109/msp.2007.163.

[89] M. Veale, R. Binns, and L. Edwards, “Algorithms that Remember: Model Inversion Attacks and Data Protection Law,” *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences*, vol. 376, no. 2133, pp. 5-6, Oct. 2018, doi: https://doi.org/10.1098/rsta.2018.0083.

[90] E. M. Power, “Developing a Culture of Privacy: a Case Study,” *IEEE Security & Privacy*, vol. 5, no. 6, pp. 281, Nov. 2007, doi: 10.1109/msp.2007.163.

[91] W. Yang et al., “Deep Learning Model Inversion Attacks and Defenses: a Comprehensive Survey,” *Artificial Intelligence Review*, vol. 58, no. 8, pp.242-243, May 2025, doi: https://doi.org/10.1007/s10462-025-11248-0.

[92] A. M. Davronbekovich, “Legal Regulation of Liability for Cyber Attacks and Data Breaches,” *International Journal of Law*, vol. 10, no. 5, 2024, pp. 113. [Online] https://www.lawjournals.org/assets/archives/2024/vol10issue5/10234.pdf.]

[93] T. Wang, Y. Zhang, and R. Jia, “Improving Robustness to Model Inversion Attacks via Mutual Information Regularization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, pp. 11666, May 2021, doi: https://doi.org/10.1609/aaai.v35i13.17387.

[94] H. Huang, “Defense Against Membership Inference Attack Applying Domain Adaptation with Addictive Noise,” *Journal of Computer and Communications*, vol. 9, no. 5, pp. 106, Jan. 2021, doi: https://doi.org/10.4236/jcc.2021.95007.

[95] Z. Zhou et al., “Model Inversion Attacks: A survey of Approaches and Countermeasures,” *arXiv (Cornell University)*, Nov. 2024, doi: https://doi.org/10.48550/arxiv.2411.10023.

[96] “OWASP Machine Learning Security Top Ten 2023 | ML03:2023 Model Inversion Attack | OWASP Foundation.” https://owasp.org/www-project-machine-learning-security-top-

10/docs/ML03_2023-Model_Inversion_Attack

[97] B. Kap, M. Aleksandrova, and T. Engel, “The Effect of Noise Level on the Accuracy of Causal Discovery Methods with Additive Noise Models,” in Luis A. Leiva et al. (eds.), *Artificial Intelligence and Machine Learning.* Cham, Springer, pp. 135.

[98] C. Aliferis and G. Simon, “Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI,” in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Health Informatics*. Cham: Springer, 2024, pp. 484-485. doi: https://doi.org/10.1007/978-3-031-39355-6_10.

[99] F. Schäfer et al., “Data-driven Business and Data Privacy: Challenges and Measures for Product-based Companies,” *Business Horizons*, vol. 66, no. 4, pp. 500, Oct. 2022, doi: https://doi.org/10.1016/j.bushor.2022.10.002.

[100] Halfond W. G., Viegas, J., and Orso, A., “A Classification of SQL-Injection Attacks and Countermeasures”. In *Proceedings of the International. Symposium on Secure Software Engineering*. 2006, pp.1. https://faculty.cc.gatech.edu/~orso/papers/halfond.viegas.orso.ISSSE06.pdf.

[101] S. Sajjadi and B. Pour, “Study of SQL Injection Attacks and Countermeasures,” *International Journal of Computer and Communication Engineering*, vol. 2, no. 5, pp.540, September 2013.

[102] F. Q. Kareem et al., “SQL Injection Attacks Prevention System Technology: review,” in *Third International Conference ACeS*, 2021, Penang, Malaysia, 24–25/08/2021, pp. 573-574. doi: https://doi.org/10.7763/IJCCE.2013.V2.244.

[103] M. Hasan, Z. Balbahaith and M. Tarique, "Detection of SQL Injection Attacks: A Machine Learning Approach," *2019 International Conference on Electrical and Computing Technologies and Applications,* Ras Al Khaimah, United Arab Emirates, 2019, pp.5, doi: https://doi.org/10.1109/ICECTA48151.2019.8959617.

[104] P. U. Alafaa, “Data privacy and Data Protection: the Right of User’s and the Responsibility of Companies in the Digital World.,” *SSRN Electronic Journal*, Jan. 2022, pp. 12, doi: https://doi.org/10.2139/ssrn.4005750.

[105] J. Mohan and V. Chidambaram, “Analyzing GDPR Compliance Through the Lens of Privacy Policy,” in *Heterogeneous Data Management*, *Polystores, and Analytics for Healthcare*, pp.82-95, 2019.

[106] S. Yaswanthra et al., “SQL Injection and Prevention,” *International Journal of Research Publication and Reviews,* vol. 5, no. 6, pp. 1311, 2024, doi: https://doi.org/10.55248/gengpi.5.0624.1438.

[107] J. Mueller, *Web Matrix Developer's Guide*. Berkeley: Apress, 2003, pp. 125.

[108] S. V. Sheta, “Challenges and Solutions in Troubleshooting Database Systems for Modern Enterprises,” *SSRN Electronic Journal*, vol. 15, no. 1, pp.53, Jan. 2024, doi: https://doi.org/10.2139/ssrn.5034358.

[109] Intersoft Consulting, “Data Protection Officer,” General Data Protection Regulation (GDPR). https://gdpr-info.eu/issues/data-protection-officer/

[110] CISSP and D. Ashbaugh, *Security Software Development: Assessing and Managing Security Risks.* Boca Raton: CRC Press, 2008, pp.13.

[111] K. Hule and R. Ranawat, “Analysis of different ORM tools for Data Access Object Tier Generation: A Brief Study,” *International Journal of Membrane Science and Technology*, vol. 10, no. 1, pp. 1289, Oct. 2023, doi: https://doi.org/10.15379/ijmst.v10i1.2842.

[112] J. Campos et al., “The Challenges of Cybersecurity Frameworks to Protect Data Required for the Development of Advanced Maintenance,” *Procedia CIRP*, vol. 47, pp. 225, 2016, doi: https://doi.org/10.1016/j.procir.2016.03.059.

[113] I. Ebert, I. Wildhaber and J. Adams-Prassl, “Big Data in the workplace: Privacy Due Diligence as a Human Rights-based Approach to Employee Privacy Protection,” *Big Data & Society*, vol. 8, no. 1, pp. 10, Jan. 2021, doi: https://doi.org/10.1177/20539517211013051.

[114] S. A. Oyetunji, “Investigating Data Protection Compliance Challenges,” *International Journal of Innovative Science and Research Technology*, pp. 2144, Sep. 2024, doi: https://doi.org/10.38124/ijisrt/ijisrt24aug1583.

Appendix-1: Abbreviation of Technical Terms

Abbreviation	***Full Form***
1NF	*First Normal Form*
2NF	*Second Normal Form*
3NF	*Third Normal Form*
ANOVA	*Analysis of Variance*
AWS	*Amazon Web Services*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CSV	*Comma Separated Values*
EMR	*Elastic MapReduce*
F-statistics	*Fisher’s F-statistics*
GBR	*Gradient Boosting Regressor*
HDFS	*Hadoop Distributed File System*
HTTPS	*Hypertext Transfer Protocol Secure*
KDE Plot	*Kernel Density Estimate Plot*
Kendall’s- “τ”	*Kendall’s Rank Correlation Coefficient*
KNN	*K-nearest Neighbour Algorithm*
MDI	*Mean Decrease Impurity*
IQR	*Interquartile Range*
ORM	*Object-Relational Mapping*
Pearson’s-“r”	*Pearson Correlation Coefficient*
P-value	*P-value for Statistical Significance*
PAAS	*Platform as a Service*
RFC	*Random Forest Classification*
SHAP	*Shapley Additive Explanations*
Spearman’s-“p”	*Spearman's Rank Correlation Coefficient*
SQL	*Structured Query Language*
SVC	*Support Vector Classification*

YARN	*Yet Another Resource Negotiator*
Z-score	*Standard Score*

Appendix-2: Data-cleaning Procedure (train dataset example)