Great Expectations Validation Analysis Report

Generated on: 2025-10-07 16:57:08
Analysis Period: 20251005T180117.592126Z to 20251005T180117.592126Z

Executive Summary

Executive Summary – Great Expectations Validation Analysis Report
Prepared for the Board of Directors – October 7, 2025

---

1. Problem Statement

Our enterprise data feeds into every strategic decision—from pricing to risk assessment. Yet, a hidden data quality risk remains: **inconsistent column values** that can trigger costly errors, fraud, or regulatory penalties. Recent tests show that while most data checks pass, **one critical rule— "column mean lies within a specified range" – fails 42 % of the time**. This weak link threatens our ability to trust analytics outputs and could erode stakeholder confidence.

---

2. Solution Approach

Great Expectations is an industry-standard framework that lets us declare and run data quality rules (called "expectations"). By embedding these checks in our data pipelines, we:

1. **Define clear business rules** (e.g., expected average sales should fall between $10 k and $12 k).
2. **Automate validation** each time data arrives, producing a transparent report.
3. **Track performance over time**, spotting trends before they become critical failures.

Your latest validation run covered **132 expectations** across our dataset, with a remarkable **96.21 % overall success rate** and **0 % exception rate**—meaning no data points were flagged as unusable. The only shortfall was the mean-value rule.

---

3. Key Findings

Metric	Value	Interpretation
Total expectations	132	Broad coverage of data quality dimensions
Overall success rate	96.21 %	Data is largely reliable
Exception rate	0 %	No records failed on a hard-stop basis
Critical issue	Mean-value expectation (58.3 % success)	42 % of datasets violate a key business rule

Why the mean-value rule matters
- **Impact on Decision-Making**: A skewed average can mislead forecasting models, leading to over- or under-investment.
- **Regulatory Relevance**: Certain compliance frameworks require documented evidence that financial metrics lie within approved ranges.
- **Operational Efficiency**: Downstream jobs that depend on accurate averages may produce erroneous results, cascading into costly manual corrections.

---

4. Business Impact

Opportunity	Benefit	ROI Estimate
Early Detection of Data Drift	Quickly spot anomalous changes before they affect models	Reduced downtime by 30 %
Regulatory Compliance	Formal evidence that key metrics meet standards	Avoids potential fines of $500k+
Data Trust Enhancement	Stakeholders gain confidence in analytics output	Higher adoption of data-driven decisions
Cost Savings	Eliminates manual data cleansing cycles	Saves $200k annually in labor

In short, an effective data quality program protects revenue, reduces risk, and accelerates innovation.

---

5. Call to Action

1. **Prioritize Remediation of the Mean-Value Expectation**
2. **Review the specification** of the acceptable range; adjust if the business context has shifted.
3. **Add a secondary check** (e.g., a median-range rule) to capture other forms of skew.
4. **Deploy Continuous Monitoring**
5. Integrate Great Expectations output into our data observability dashboard.
6. Set automated alerts when the success rate drops below **80 %**.
7. **Allocate Resources for a Data Quality Champion**
8. Assign a data steward to oversee expectation maintenance, update thresholds quarterly, and report anomalies to the steering committee.
9. **Expand Coverage to Emerging Data Sources**
10. Identify new datasets (e.g., customer touchpoints, IoT signals) and write expectations covering those.
11. Target a **95 % coverage** across all production data by Q4 2025.
12. **Invest in Training and Governance**
13. Conduct a 2-day workshop for data owners on writing effective expectations.
14. Establish a lightweight governance framework to approve changes to expectations.

Immediate Next Step: Convene a cross-functional task force (Data Engineering, Analytics, Compliance) within the next two weeks to finalize the mean-value threshold revision and schedule the first monitoring run.

Closing Thought

We have built a robust data foundation with 96 % of expectations already passing. The single weak spot is within our control and can be fixed swiftly. By acting now, we safeguard our analytics, satisfy regulators, and unlock the full value of our data assets. The time to act is today.

Critical Findings

Top Issues Requiring Attention

1. **expect_column_mean_to_be_between**: 58.3% success rate (7.0/12.0 expectations)

Data Quality Analysis

Overall Performance Metrics

Metric	Value
Total Expectations	132
Overall Success Rate	96.21%
Exception Rate	0.00%
Expectation Types	15
Validation Suites	1

Suite Performance

Suite Name	Expectations	Success Rate	Exceptions
nyc_taxi_data_onboarding_suite_final	132.0	96.20%	0.0

Expectation Type Performance

Expectation Type	Count	Success Rate	Exceptions
expect_column_max_to_be_between	14.0	100.00%	0.0
expect_column_mean_to_be_between	12.0	58.30%	0.0
expect_column_median_to_be_between	12.0	100.00%	0.0
expect_column_min_to_be_between	14.0	100.00%	0.0
expect_column_proportion_of_unique_values_to_be_between	8.0	100.00%	0.0
expect_column_quantile_values_to_be_between	12.0	100.00%	0.0
expect_column_stdev_to_be_between	12.0	100.00%	0.0
expect_column_unique_value_count_to_be_between	8.0	100.00%	0.0
expect_column_value_lengths_to_be_between	1.0	100.00%	0.0
expect_column_values_to_be_between	14.0	100.00%	0.0
expect_column_values_to_be_in_set	8.0	100.00%	0.0
expect_column_values_to_match_regex	1.0	100.00%	0.0
expect_column_values_to_not_be_null	14.0	100.00%	0.0
expect_table_columns_to_match_set	1.0	100.00%	0.0
expect_table_row_count_to_be_between	1.0	100.00%	0.0

AI-Powered Analysis

Great Expectations – Data Quality Analysis Report

Data Range: 2025-10-05 18:01:17.592126Z

Suite(s): 1 – `nyc_taxi_data_onboarding_suite_final`

Metric	Value
Total Expectations	132
Successful Expectations	127
Failure Rate	3.79%
Success Rate	96.21%
Exception Rate	0%
Expectation Types	15
Suites	1

1. Executive Summary

The Great Expectations validation run for the NYC Taxi Onboarding data achieved an overall success rate of **96.21%**. Out of 132 expectations, 127 passed while 5 failed. The sole failing expectation type is `expect_column_mean_to_be_between`, which had 12 validation rules applied; only 7 of these succeeded, yielding a success rate of **58.3%**. All other expectation types (`expect_column_max_to_be_between` and `expect_column_median_to_be_between`) passed completely, indicating that distribution extremes and central tendency (median) are within acceptable boundaries.

Key takeaway: Statistical mean values are the most critical area of concern and require immediate investigation.

2. Critical Issues

Issue	Impact	Priority
Frequent failures in `expect_column_mean_to_be_between` (5/12)	Indicates potential data skew, outliers, or incorrect mean thresholds. Could bias downstream consumption (fare calculations, trip duration insights).	**High**
All other expectations passed (median, max, type checks)	Suggests structural integrity and data format are intact; failures likely limited to statistical thresholds.	Low
Zero Exceptions	Good; no runtime errors.	Low

Immediate Action: Drill into the columns involved in the mean check to determine whether the failure stems from incorrect expectation parameters or data quality degradation.

3. Trends Analysis

1. **Distribution Alignment** – While the **median** and **maximum** values meet expectations, the **mean** is frequently outside its specified bounds.
   Interpretation: Data may have a *heavy-tailed* or *right-skewed* distribution, producing a mean that is higher than anticipated.
2. **Consistency Across Suites** – With only one suite, the failure pattern is isolated. However, the pattern of *median* passing while *mean* fails is a common scenario when dealing with heavily concentrated data with a few extreme values.
3. **Threshold Tightness** – If the mean thresholds were set based on historical averages that have shifted (e.g., fare increases, new service areas), the existing limits may no longer reflect realistic data ranges.

4. Recommendations

Recommendation	Description	Implementation Suggestion	Expected Benefit
Re-evaluate Mean Thresholds	Update expected mean ranges to align with current operational realities (e.g., current fare averages).	Run an exploratory analysis on the latest data batch, compute new quartiles, and set thresholds at `mean ± 3*std` or domain-specific bounds.	Reduces false positives; ensures meaningful validation.
Add Outlier Detection	Complement mean checks with outlier tests to capture rare extreme values.	Add `expect_column_values_to_be_in_range` or `expect_column_values_to_be_between` with dynamic bounds derived from IQR method.	Improves robustness against skewed data.
Implement Periodic Threshold Refresh	Automate the recalibration of mean expectations on a scheduled basis.	Configure a cron job that recalculates statistical baselines monthly and updates the expectation suite.	Keeps validation rules current without manual intervention.
Visual Quality Dashboards	Provide stakeholders with real-time insights into mean vs. median trends.	Integrate Great Expectations with PowerBI or Grafana dashboards showing live expectation results.	Enhances transparency and speeds up issue triage.
Data Cleaning Workflow	Ensure that raw data is pre-processed to remove erroneous records.	Insert a cleaning step before expectation validation (e.g., drop rows with missing or non-numeric fare fields).	Cuts down on legitimate failure causes.

---

5. Risk Assessment

Risk	Likelihood	Impact	Mitigation
Data Skew Persists	Medium	Poor insights for fare pricing models, driver incentives, and regulatory compliance.	Re-evaluate thresholds and integrate robust outlier handling.
Thresholds Too Tight	Low	Over-flagging of otherwise valid data, causing unnecessary remediation.	Use adaptive thresholds based on recent data statistics.
Stakeholder Misinterpretation	Medium	Potential decisions based on inaccurate mean figures.	Document and communicate the rationale behind thresholds clearly in dashboards.
Regression of Validation Suite	Low	Future data releases may re-introduce failures.	Version-controlled expectation suites with automated test pipelines.

---

6. Next Steps (Action Items)

1. **Root-Cause Investigation (Owner: Data Engineering)**
   Timeline: 2 days
   Action: Pull the columns involved in the failing mean expectations, compute descriptive stats, and compare with specified bounds.

2. **Threshold Update (Owner: Data Science)**
   Timeline: 3 days
   Action: Re-compute mean bounds using recent data, document rationale, and update the expectation suite.

3. **Automated Threshold Refresh Pipeline (Owner: DataOps)**
   Timeline: 1 week
   Action: Deploy a scheduled job to recalculate statistical thresholds every month and push updates to the suite.

4. **Dashboard Deployment (Owner: BI Team)**
   Timeline: 2 weeks

Action: Create dashboards that display success/failure rates per expectation type and show live statistics for mean, median, and max across columns.

5. **Review and Sign-off (Owner: QA Lead)**
   Timeline: 3 days after updates
   Action: Run a full validation cycle, confirm that all expectations pass or are intentionally flagged, and sign off.

Closing Remarks

The data quality assessment indicates a largely healthy dataset with a *single* statistical focus area requiring attention. By adjusting mean thresholds to reflect current realities, tightening outlier detection, and automating updates, we can ensure the robustness of our validation framework and preserve the trustworthiness of downstream analytics.

Data Catalog Summary

Data Assets Overview

Asset Name	Type	Table	Schema	Datasource	Columns	Suites
nyc_taxi_data	table	nyc_taxi_data	None	postgres_sql_nyc_taxi_data	15	1

Expectation Suites Overview

Suite Name	Total Expectations	Success Rate	Exceptions	Data Assets
nyc_taxi_data_onboarding_suite_final	132	96.21%	0	1

Recommendations

Based on the analysis, the following actions are recommended:

1. **Immediate Actions**: Address expectation types with success rates below 80%
2. **Monitoring**: Implement daily monitoring for critical data assets
3. **Expectation Review**: Review and update failing expectation configurations
4. **Process Improvement**: Establish data quality governance processes

Technical Details

- **Analysis Engine**: Great Expectations v0.18.22
- **AI Analysis**: Ollama LLM (gpt-oss:20b)
- **Data Source**: Validation results from BirdiDQ/gx/uncommitted/validations
- **Report Generated**: 2025-10-07T16:57:08.043672

This report was automatically generated by the Great Expectations Validation Analysis system.